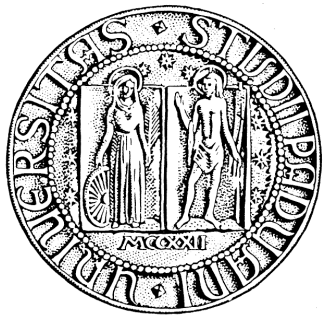
UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

Corso di Laurea Magistrale in Scienze Statistiche



Tesi di Laurea

UN MODELLO BAYESIANO NON PARAMETRICO PER L'ANALISI DEL TRAFFICO TELEFONICO

Relatore: Prof. Bruno Scarpa

Laureanda: Francesca Penzo

ANNO ACCADEMICO 2010/2011

Alla mia famiglia

Indice

Elenco delle figure

Elenco delle tabelle

1 Introduzione

Grazie alla diffusione delle nuove tecnologie informatiche, nel corso degli ultimi anni è mutato profondamente il modo in cui aziende e clienti interagiscono tra loro. Il successo di un'impresa non dipende più esclusivamente dalle specifiche attività della catena del valore, ma anche, e in misura sempre maggiore, dalla capacità di istituire relazioni stabili con i clienti. In questo contesto assume un ruolo essenziale la formulazione di una strategia di *Customer Relationship Management* (CRM), cioè una strategia aziendale focalizzata sull'interpretazione e il soddisfacimento delle esigenze dei clienti (Farinet e Ploncher, 2002).

Uno dei problemi di CRM che spesso le imprese si trovano ad affrontare riguarda la fidelizzazione dei clienti: le aziende caratterizzate da un rapporto continuativo con i propri clienti, quali società di servizi, telecomunicazioni, banche, ecc. operano spesso in mercati saturi e caratterizzati da elevata concorrenza, perciò il costo di acquisizione di nuovi clienti è notevole ed essi diventano profittevoli per l'azienda solo dopo un lungo periodo. Quindi è fondamentale riuscire a trattenere i propri clienti per recuperare i costi sostenuti, evitando che essi abbandonino l'azienda per passare ad un concorrente (tale fenomeno è indicato con il termine inglese *churn*). Per un'azienda che opera in questo contesto è importante dotarsi di modelli statistici che permettano di individuare tempestivamente i clienti a rischio di abbandono per poter attuare azioni di trattenimento.

Lo scopo di questa tesi consiste nella previsione del tasso di *churn* per una società di telecomunicazioni, utilizzando i dati sui clienti già disponibili all'interno dell'azienda stessa. Possiamo suddividere tali dati in due tipologie: i dati di tipo "statico", cioè quelle informazioni sul cliente, raccolte all'atto della sottoscrizione del contratto, che generalmente consideriamo invarianti nel tempo nel periodo di riferimento (come le caratteristiche socio-demografiche). Consideriamo come "dinamiche", invece, le informazione relative al traffico telefonico: vengono rilevate periodicamente (spesso a cadenza mensile) e possono variare nel tempo.

Per raggiungere il nostro obiettivo potremmo applicare metodi di classificazione o strumenti di *data mining* (Azzalini e Scarpa, 2004), ma generalmente queste procedure trattano tutti i dati come se fossero statici, mentre l'informazione relativa all'andamento nel tempo del traffico telefonico potrebbe fornire elementi utili per caratterizzare in maniera più precisa il comportamento del cliente.

La recente letteratura in ambito biostatistico propone degli strumenti di tipo bayesiano non parametrico per la modellazione congiunta di una traiettoria (trattata come dato funzionale) e una variabile risposta, si vedano ad esempio Dunson, Herring e Siega-Riz (2008) e Bigelow e Dunson (2009).

In questa tesi applicheremo l'approccio sviluppato in ambito biostatistico al contesto della modellazione del *churn*.

Il Capitolo 2 è dedicato alle analisi esplorative condotte sull'insieme di dati a disposizione. Il Capitolo 3 fornisce una breve descrizione dell'analisi bayesiana in ambito non parametrico. Nel Capitolo 4 presentiamo il modello sviluppato per l'analisi. Nel Capitolo 5 applichiamo tale modello all'insieme di dati a disposizione, considerando solo le variabili "dinamiche" per prevedere il *churn*. Infine, nel Capitolo 6 estendiamo il modello per includere anche le variabili "statiche" e confrontiamo i risultati con quelli ottenuti nel Capitolo precedente.

Tutte le analisi sono state condotte utilizzando il software statistico R[1].

[1] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

2 Dati di telefonia

2.1. Descrizione dell'insieme dei dati

I dati a disposizione[1] costituiscono un campione di 3000 unità estratto casualmente dal DWH (*Data WareHouse*) di una compagnia telefonica europea e si riferiscono a clienti con contratto post-pagato attivo per dieci mesi consecutivi. Per ogni cliente sono state osservate le seguenti variabili: il numero di chiamate in uscita effettuate in ciascuno dei dieci mesi considerati e lo stato del contratto (ancora attivo oppure disattivato) due mesi dopo l'ultimo mese in cui sono disponibili i dati sul traffico telefonico. Nella modellazione del *churn* è utile non considerare il traffico dei periodi immediatamente precedenti alla rilevazione dello stato del cliente, per poter individuare i clienti a rischio abbandono con uno o due mesi di anticipo sull'abbandono effettivo. In questo modo l'azienda ha il tempo necessario per attuare azioni di *retention* nei confronti del cliente.

In una prima analisi delle serie storiche è emerso un andamento anomalo nel decimo mese rispetto ai mesi precedenti, di cui non si conoscono le cause. Per questo motivo, nelle analisi successive consideriamo le serie storiche fino al nono mese disponibile.

Gli utenti disattivati sono 443 e costituiscono circa il 15% del campione. Analizziamo l'andamento delle serie storiche nel tempo per i due gruppi (utenti attivi e disattivati): come si può notare dalla Figura 2.1, il numero mensile di chiamate in uscita si concentra su valori inferiori a 400 telefonate. La linea rossa rappresenta l'andamento medio delle telefonate per ciascun gruppo. La variabile assume complessivamente un ampio *range* di valori: il numero minimo di telefonate per ogni mese è zero, mentre il massimo varia tra 993 telefonate per il secondo mese e 1340 per il nono.

[1] Si tratta dei dati di clienti di telefonia utilizzati per alcune analisi nel testo Azzalini e Scarpa (2004), reperibili all'indirizzo web: http://azzalini.stat.unipd.it/Libro-DM/

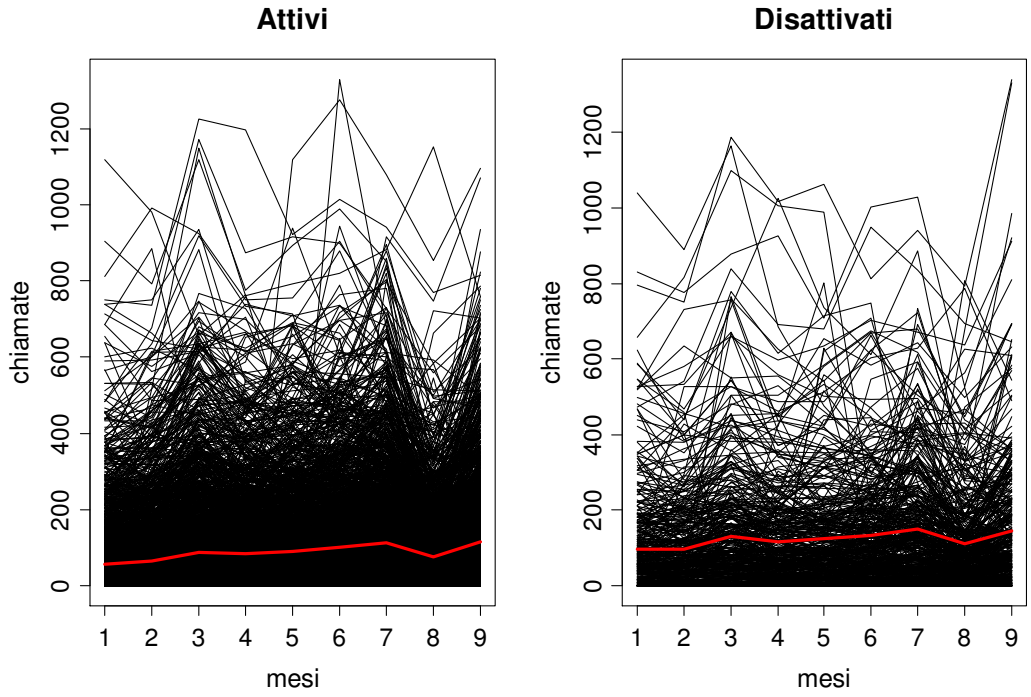


Figura 2.1 Rappresentazione del numero di chiamate in uscita rispetto al tempo

L'elevata numerosità del campione non permette un'agevole lettura del grafico, quindi rappresentiamo in Figura 2.2 esclusivamente la media complessiva del campione e le medie dei due sottogruppi (attivi e disattivati).

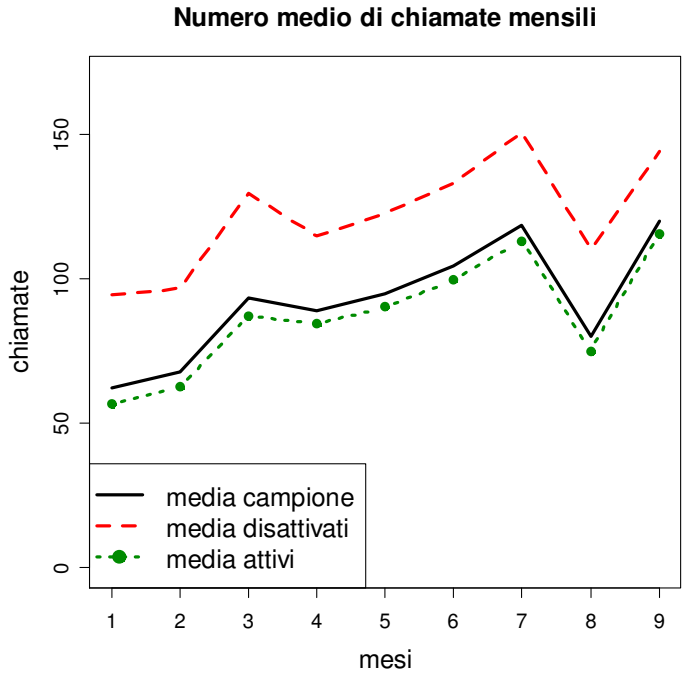


Figura 2.2 Grafico della media complessiva del campione e delle medie per i due gruppi di clienti (attivi e disattivati)

La media degli attivi è prossima a quella del campione, come ci si aspetta, dato che gli utenti attivi sono l'85% del totale. Per entrambe le traiettorie medie si nota un andamento decrescente in corrispondenza dell'ottavo mese; risulta difficile fornire un'interpretazione a tale andamento, poiché non conosciamo il mese né l'anno a cui corrisponde.

Per quanto riguarda l'andamento medio dei clienti disattivati, si nota che per ogni mese è chiaramente distinguibile da quello degli attivi ed è più elevato. Questo particolare può risultare insolito, perché potrebbe significare che, mediamente, gli utenti che si disattiveranno effettuano più chiamate rispetto agli altri. Esaminando più in dettaglio le serie storiche (si vedano Tabella 2.1 e Figura 2.3[2]), si nota come la distribuzione del numero di chiamate per gli utenti attivi sia concentrata su valori più bassi rispetto a quella degli utenti disattivati. Infatti, come si vede dalla Figura 2.4, i clienti attivi hanno un percentuale di valori nulli per il numero di chiamate sempre inferiori rispetto agli utenti attivi, tranne negli ultimi due mesi.

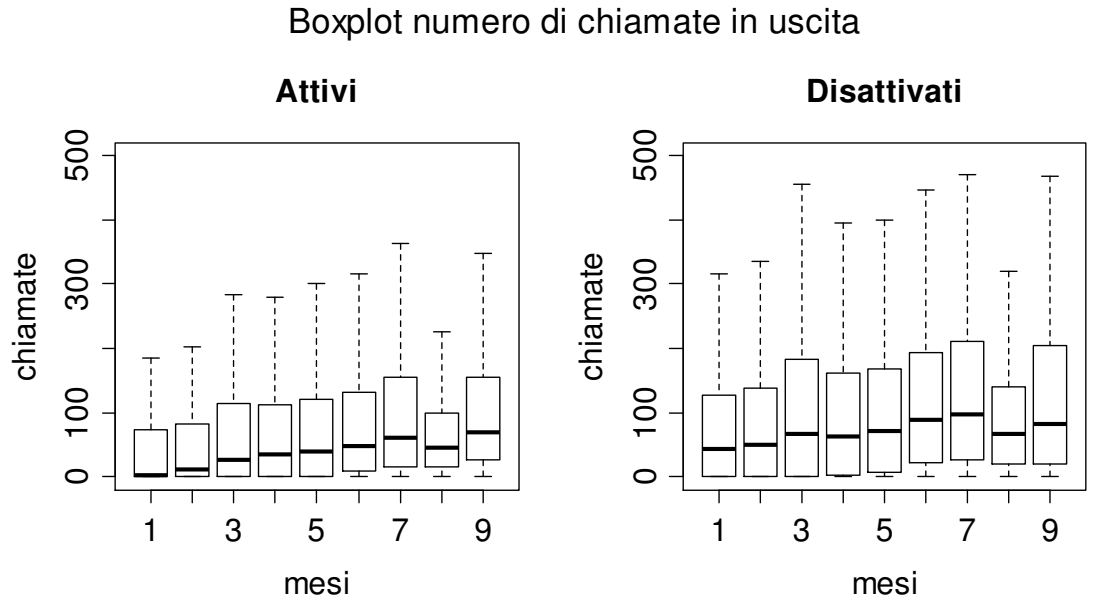


Figura 2.3 Boxplot del numero di chiamate in uscita per i due gruppi (utenti attivi e disattivati)

[2] Nella rappresentazione abbiamo escluso i valori esterni ai "baffi" delle scatole, per maggior chiarezza, tuttavia, tali osservazioni non verranno escluse nelle successive analisi.

Attivi							
Mese	Media	Deviazione standard	Minimo	Primo quartile	Mediana	Terzo quartile	Massimo
1	56,51	102,13	0,00	0,00	3,00	74,00	1119,00
2	62,66	109,72	0,00	0,00	12,00	81,00	993,00
3	87,12	140,90	0,00	0,00	27,00	114,00	1225,00
4	84,28	124,75	0,00	0,00	35,00	112,00	1198,00
5	90,17	129,49	0,00	0,00	40,00	121,00	1120,00
6	99,51	137,83	0,00	9,00	48,00	132,00	1330,00
7	113,00	143,59	0,00	16,00	61,00	155,00	1081,00
8	74,87	94,04	0,00	15,00	45,00	99,00	1152,00
9	115,60	135,11	0,00	26,00	69,00	155,00	1097,00

Disattivati							
Mese	Media	Deviazione standard	Minimo	Primo quartile	Mediana	Terzo quartile	Massimo
1	94,48	141,39	0,00	0,00	44,00	126,50	1041,00
2	96,75	137,20	0,00	0,00	50,00	137,00	890,00
3	129,70	181,67	0,00	0,00	66,00	182,00	1187,00
4	114,90	155,85	0,00	2,00	62,00	161,50	1026,00
5	122,50	160,51	0,00	8,00	71,00	168,50	1062,00
6	132,90	156,87	0,00	22,50	88,00	194,00	1002,00
7	150,50	170,94	0,00	27,00	98,00	209,50	1029,00
8	110,20	135,22	0,00	20,50	68,00	140,00	806,00
9	144,10	181,61	0,00	20,00	81,00	205,00	1340,00

Tabella 2.1 Sintesi della distribuzione del numero di chiamate in uscita per ciascun mese disponibile

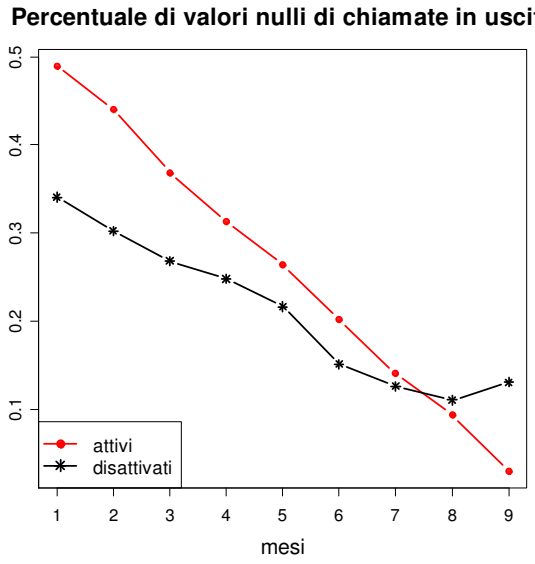


Figura 2.4 Percentuale mensile di valori nulli per i gruppi dei clienti attivi e dei disattivati

2.2. Dai dati al modello

Come abbiamo già accennato nel capitolo introduttivo, gli approcci classici di *data mining* come alberi di classificazione, reti neurali, ecc. non sono adatti a modellare dati dinamici, perché non tengono conto della struttura temporale degli stessi. Per questo motivo si è deciso di procedere nell'analisi del traffico telefonico considerando il numero di chiamate nei mesi a disposizione come un dato funzionale.

Per trattare la relazione tra predittori funzionali e una risposta scalare, una strategia comune consiste nella definizione di alcune caratteristiche che riassumono la funzione, come il tasso di variazione o il valore medio negli istanti temporali considerati, ecc. Questi indicatori possono essere inclusi come variabili esplicative in un modello lineare generalizzato (GLM) per la variabile risposta, tuttavia, spesso non è chiaro quali siano gli indicatori migliori per riassumere la funzione, e sceglierne molti può causare problemi di multicollinearità nel modello. Inoltre, sorgono delle difficoltà se le misura del dato funzionale sono effettuate a istanti differenti per ciascun individuo. In letteratura sono state proposte delle modifiche ai modelli classici per trattare la relazione tra predittori funzionali e una risposta scalare: James (2002) ha esteso i GLM per includere predittori funzionali; James e Silverman (2005) hanno proposto per lo stesso scopo una classe più generale di modelli che estende GLM, regressione *projection pursuit* e modelli additivi generalizzati.

In alternativa, potremmo considerare una procedura a due stadi: prima stimiamo un modello per classificare le sole traiettorie, considerate come dati longitudinali, successivamente includiamo gli indicatori di cluster in un modello di regressione per la variabile risposta. Un tale approccio, tuttavia, non permette ai valori della risposta di essere informativi nella procedura di *clustering* e quindi può causare delle distorsioni.

Una strategia alternativa consiste nell'utilizzo di una metodologia di *clustering* flessibile in cui i gruppi sono definiti sia dai predittori funzionali sia dal livello della variabile risposta. Bigelow e Dunson (2009) hanno esteso l'approccio bayesiano semiparametrico sviluppato per il *clustering* flessibile alla modellazione congiunta di dati funzionali con una variabile risposta: hanno utilizzato un modello *multivariate adaptive spline* per descrivere i predittori funzionali e un modello

lineare generalizzato con intercetta casuale per descrivere la variabile risposta. Specificando come *a priori* un Processo di Dirichlet (Ferguson, 1973, 1974) per la distribuzione dell'intercetta casuale congiuntamente ai coefficienti delle basi di funzioni, hanno ottenuto una procedura che raggruppa le traiettorie secondo la forma e secondo i parametri del modello per la risposta.

Tale approccio è utile per l'analisi in contesti come quello preso in esame in questa tesi, in cui sono disponibili poche informazioni a priori circa la forma delle funzioni casuali oggetto di studio e si desidera un modello caratterizzato da elevata flessibilità. Per questo motivo applicheremo l'analisi bayesiana non parametrica alla modellazione del *churn*.

3 Analisi bayesiana non parametrica

Lo scopo di questo capitolo è presentare una breve descrizione dell'analisi bayesiana in ambito non parametrico. Per una trattazione più ampia si rimanda a Dunson (2010) e, più in generale, a Hjort (2010).

3.1. Il Processo di Dirichlet nella modellazione di dati gerarchici

La modellazione gerarchica è diventata uno strumento comune nello studio della dipendenza per dati di tipo longitudinale. Un semplice modello gerarchico ha la seguente forma:

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \mu_i &\sim P \end{aligned} \tag{3.1}$$

dove y_{ij} è la j-ma osservazione relativa all'individuo o al gruppo i, con $i = 1, \dots, n$ e $j = 1, \dots, n_i$, μ_i è la media specifica per ciascun individuo, ε_{ij} è un residuo relativo alla singola osservazione, σ^2 è la varianza tra le osservazioni relative allo stesso individuo, P è la distribuzione della media μ_i.

La specificazione parametrica usuale per il modello (3.1) prevede $\mu_i = \mu + b_i$, dove μ è la media complessiva e b_i un effetto casuale per il soggetto i. L'eterogeneità tra gli individui è caratterizzata dalla distribuzione dell'effetto casuale e solitamente si pone $b_i \sim \mathcal{N}(0, \psi)$. Da queste assunzioni segue che $P \sim \mathcal{N}(\mu, \psi)$. Tuttavia, l'assunzione di normalità su P potrebbe non essere adeguata, in particolare perché tale distribuzione prevede delle code poco pesanti e quindi non permette che gli individui differiscano molto tra loro, ma anche per la forma simmetrica e unimodale, che spesso non ha giustificazioni a priori per essere assunta. Lee e Thompson (2008) hanno descritto alcuni approcci bayesiani parametrici per la modellazione flessibile della distribuzione degli effetti casuali, tuttavia i modelli parametrici, come per esempio estensioni della distribuzione t, sono comunque restrittivi e non permettono la presenza di più mode, necessaria se nei dati sono presenti sub-popolazioni latenti.

Per aumentare la flessibilità della rappresentazione di P è possibile ricorrere a modelli Bayesiani di tipo non parametrico. Tali modelli centrano la distribuzione a priori su un modello parametrico di base e includono un numero infinito di parametri per permettere un elevato grado di flessibilità. Quindi, in assenza di conoscenza parametriche su P, è possibile scegliere una distribuzione a priori che abbia come supporto l'insieme della distribuzioni sull'asse reale, in questo modo la distribuzione a priori corrisponde ad una distribuzione sulle distribuzioni.

Bush e MacEachern (1996) hanno proposto di trattare questo problema scegliendo un Processo di Dirichlet (DP) per la distribuzione a priori di P (Ferguson, 1973, 1974). Si assume, quindi, $P \sim DP(\alpha P_0)$, dove $\alpha > 0$ è un parametro di concentrazione che caratterizza la precisione della distribuzione a priori, mentre P_0 è la distribuzione base su $\Re$, cioè la migliore previsione di P a priori. Spesso si sceglie $P_0 \sim \mathcal{N}(\mu_0, \psi_0)$.

Per meglio comprendere cosa implica la scelta di una *a priori* DP per P, consideriamo la rappresentazione *stick-breaking* del DP definita da Sethuraman (1994): assumere $P \sim DP(\alpha P_0)$ equivale ad assumere

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \overset{iid}{\sim} H_0 \tag{3.2}$$

dove $\{\pi_h\}_{h=1}^{\infty}$ sono i pesi di probabilità definiti mediante un processo *stick-breaking*: $\pi_h = V_h \prod_{l=1}^{h-1}(1 - V_l)$ con $V_h \sim Beta(1, \alpha)$, e δ_θ denota la misura di probabilità di Dirac sull'atomo θ.

La denominazione "*stick-breaking*" per tale processo deriva dal fatto che, supponendo di iniziare il processo con un segmento di probabilità unitaria, V_1è la porzione di segmento "spezzata" ed assegnata a θ_1, V_2 è la porzione della rimanente parte $1 - V_1$ assegnata a θ_2 e così via (si veda Figura 3.1 a pagina seguente).

Per valori di α prossimi a zero, $V_1 \approx 1$ e tutta la probabilità viene assegnata ad un singolo atomo, per valori piccoli di α, per esempio $\alpha = 1$, la maggior parte della probabilità è assegnata ai primi atomi, mentre per valori grandi di α ad ogni atomo è assegnato un peso di probabilità molto piccolo, così che P risulta simile a P_0. La scelta usuale è fissare $\alpha = 1$, in alternativa è possibile specificare per α una distribuzione iper a priori (Escobar e West, 1995).

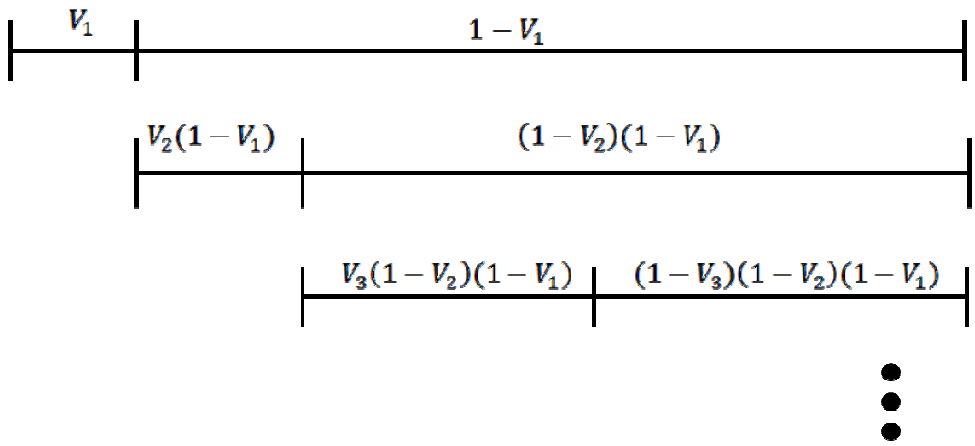


Figura 3.1 Rappresentazione del processo stick-breaking

Un'importante implicazione di (3.2) è la natura quasi certamente discreta di P (Sethuraman, 1994). Questo crea dei vincoli tra le medie $\mu_i, i = 1, \dots, n$ e la configurazione di tali vincoli definisce cluster in cui il valore dell'effetto casuale è lo stesso per tutti gli individui che appartengono al medesimo cluster. Indicando con $S_i = k$ l'appartenenza dell'individuo i al cluster k, si ha che $\mu_{S_i} = \theta^*_{S_i}$ per $i = 1, \dots, n$, dove θ^*_k indica il valore dell'effetto casuale per tutti i soggetti del cluster k (è opportuno sottolineare la distinzione tra θ^*_k, il k-mo cluster rappresentato nel campione di n individui, e θ_k, k-mo degli infiniti atomi nella rappresentazione *stick-breaking*).

Il *clustering* mediante DP presenta alcuni vantaggi rispetto ad altre procedure: evita di assumere che gli individui siano raggruppati in un fissato numero di cluster, poiché, come è chiaro dalla rappresentazione *stick-breaking*, il DP assume la presenza di infiniti cluster nella popolazione complessiva e solo un numero ignoto di tali cluster è osservato in un campione finito di n individui. Quando un nuovo individuo si aggiunge al campione, esiste una probabilità non nulla, $\alpha/(\alpha + n)$, che sia assegnato ad un cluster non ancora rappresentato nel campione. Tuttavia, tale probabilità diminuisce all'aumentare di n e ciò costituisce una delle penalità per la complessità del modello, sia esplicite sia implicite, che evitano di assegnare ciascun individuo ad un gruppo diverso per ottenere una verosimiglianza maggiore.

La flessibilità di tale approccio permette di non fissare a priori il numero di gruppi, tuttavia causa delle difficoltà nell'interpretazione dei risultati, perché il numero di cluster e la loro composizione varia tra le iterazioni dell'algoritmo

necessario per il calcolo della distribuzione a posteriori dei parametri. Questo problema è noto in letteratura con il nome di *label switching* e sarà discusso più dettagliatamente nel Paragrafo 4.4.

3.2. Mistura di DP

Il modello gerarchico in (3.1) può essere riscritto come un modello di mistura:

$$y_i = \int \mathcal{N}(y_i; \mu_i, \sigma^2) dP(\mu_i) \tag{3.3}$$

Per evitare di trattare l'infinito numero di parametri che caratterizzano P (si veda l'espressione (3.2)), operiamo una marginalizzazione rispetto a P nell'espressione (3.3) e consideriamo lo schema urne di Pólya di Blackwell e MacQueen (1973). In particolare otteniamo $(\mu_1, \dots, \mu_n) \sim PU(\alpha P_0)$. La distribuzione a priori per l'elemento μ_i condizionata agli elementi $\boldsymbol{\mu}^{(i)} = (\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_n)$ è

$$p\left(\mu_i \middle| \boldsymbol{\mu}^{(i)}, \alpha\right) \propto \left(\frac{\alpha}{\alpha + n - 1}\right) P_0 + \left(\frac{1}{\alpha + n - 1}\right) \sum_{i' \neq i} \delta_{\mu_{i'}} \tag{3.4}$$

Poiché molti degli elementi in $\boldsymbol{\mu}^{(i)}$ sono già raggruppati in cluster, possiamo riscrivere l'espressione (3.4) nel modo seguente:

$$p\left(\mu_i \middle| \boldsymbol{\theta}^{(i)}, \alpha\right) \propto \left(\frac{\alpha}{\alpha + n - 1}\right) P_0 + \left(\frac{1}{\alpha + n - 1}\right) \sum_{j=1}^{k^{(i)}} n_j^{(i)} \delta_{\theta_j^{(i)}} \tag{3.5}$$

dove $\boldsymbol{\theta}^{(i)} = \left(\theta_1^{(i)}, \dots, \theta_{k^{(i)}}^{(i)}\right)'$ rappresenta i $k^{(i)}$ valori distinti di $\boldsymbol{\mu}^{(i)}$ e $n_j^{(i)}$ è il numero di elementi di $\boldsymbol{\mu}^{(i)}$ allocati nel cluster j.

L'integrazione rispetto alla distribuzione infinito-dimensionale P induce sull'effetto casuale dell'i-mo individuo una distribuzione a priori, condizionata agli effetti casuali degli altri individui, costituita da una mistura della distribuzione di base e una distribuzione uniforme discreta sui valori degli altri soggetti. Questo processo mostra in maniera più esplicita il *clustering* che deriva dall'assumere una *a priori* DP per P: gli individui vengono raggruppati in cluster, e occasionalmente viene assegnato un soggetto ad un nuovo cluster con effetto casuale estratto dalla distribuzione di base.

Il parametro α del DP controlla il numero di componenti della mistura:
se $\alpha \to 0$ si ha un'unica componente nella mistura, e quindi stimiamo un modello senza eterogeneità tra individui: si ha $\mu_i = \mu$ e

$$y_i|\mu \sim_{iid} \mathcal{N}(y_i; \mu, \sigma^2) \quad \mu \sim P_0$$

se $\alpha \to \infty$ si ha $\mu_i \sim P$ e quindi una componente per ciascuna osservazione. In questo caso la stima del modello non parametrico equivale alla stima di un modello gerarchico parametrico.

Quindi il parametro α controlla anche quanti elementi distinti ci sono nel campione $\mu_1, \dots, \mu_n$. In questo modo non è necessario specificare a priori il numero di cluster, poiché viene trattato come una variabile casuale. Sia k il numero di cluster presenti nel campione, allora:

$$P(k = m|\alpha, n) = c_n(m) n!\, \alpha^h \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad m = 1, \dots, n \tag{3.6}$$

dove $c_n(m) = P(K = m|\alpha = 1, n)$ (Antoniak, 1974).

Il valore atteso a priori del numero di cluster è proporzionale a $\alpha \log n$, in questo modo il numero di gruppi aumenta lentamente con il numero di unità del campione ad un tasso determinato da α. Il parametro α riveste, quindi, un ruolo chiave nella determinazione del numero di gruppi.

3.3. Calcolo della distribuzione a posteriori

Dopo aver specificato le distribuzioni a priori, è necessario considerare come aggiornare queste distribuzioni con le informazioni derivanti dai dati per ottenere la distribuzioni a posteriori.

Anche per modelli gerarchici semplici, come il modello specificato in (3.1), la distribuzione a posteriori è spesso non trattabile analiticamente e per questo è necessario far ricorso ad approssimazioni numeriche: l'approccio standard è costituito dagli algoritmi Markov Chain Monte Carlo (MCMC), di cui sono state proposte varie alternative. Le principali tipologie di algoritmi MCMC includono il *collapsed Gibbs sampler* (MacEachern, 1994) e il *blocked Gibbs sampler* (Ishwaran e James, 2001). Quest'ultimo si basa su un'approssimazione finita di P attraverso il troncamento della rappresentazione *stick-breaking* in (3.2) proposto originariamente da Muliere e Tardella (1998). Poiché i pesi di probabilità assegnati

agli atomi tendono a decrescere rapidamente all'aumentare dell'indice h, è ragionevole sostituire la somma infinita con una somma dei primi N termini; ciò può essere realizzato semplicemente ponendo $V_N = 1$.

L'algoritmo *collapsed Gibbs sampler*, invece, evita di fissare la soglia N per il numero di cluster, mantenendo il modello di mistura con un numero infinito di componenti. Per il calcolo della distribuzione a posteriori dei parametri si fa ricorso allo schema ad urne di Pólya (si veda Sezione 3.2). Ora descriviamo brevemente tale schema, utilizzando la metafora del ristorante cinese, nel Capitolo 4 lo presenteremo in maggior dettaglio, poiché utilizzeremo tale algoritmo nell'analisi.

Il *Chinese Restaurant Process*, CRP, (Pitman, 1996), descrive la probabilità marginale del DP in termini di una partizione casuale ottenuta da una sequenza di clienti che siedono ai tavoli di un ristorante. I clienti arrivano in successione ad un ristorante che ha infiniti tavoli, a ciascuno dei quali può sedere un numero infinito di persone. Il primo cliente che arriva siede al primo tavolo; quando arrivano altri clienti, questi possono essere assegnati ad un tavolo occupato oppure ad uno libero. Per l'i-mo cliente, la probabilità di essere assegnato ad un tavolo libero è $\alpha/(\alpha + i - 1)$, mentre la probabilità di essere assegnato al j-mo tavolo già occupato è $n_{ij}/(\alpha + i - 1)$, dove n_{ij} indica il numero di clienti già seduti al tavolo j quando arriva il cliente i. Il CRP alloca i clienti ai tavoli secondo lo schema ad urne di Pólya di $DP(\alpha P_0)$, dove gli individui seduti al tavolo j condividono un piatto θ_j^* estratto dalla distribuzione P_0. Nella metafora, ogni cliente è un'unità del campione di dati, tavoli rappresentano i cluster e il piatto è il valore del parametro specifico per ciascun cluster (si veda Figura 3.2).

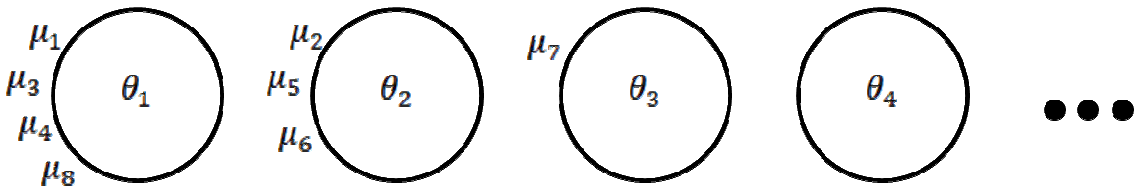


Figura 3.2 Rappresentazione del CRP

Da questo schema si nota che pochi tavoli attraggono la maggior parte dei clienti: infatti, un cliente siede ad un tavolo con probabilità proporzionale al numero di clienti già seduti a quel tavolo (mentre siede ad un tavolo libero con

probabilità proporzionale al parametro di concentrazione α), quindi i clienti tendono a sedersi nei tavoli più "popolari" e, poiché questi ultimi attraggono la maggior parte dei nuovi clienti, diventano sempre più "popolari".

3.4. Analisi di dati funzionali

Sono molti i campi in cui l'interesse è rivolto studio della variabilità nelle funzioni casuali: da quello biologico, come negli studi sulle traiettorie ormonali condotti da Scarpa e Dunson (2009) e da Bigelow e Dunson (2009), a quello economico, con lo studio dell'andamento degli indici azionari, a quello aziendale, con lo studio dell'andamento del traffico telefonico, ecc. I metodi di analisi di dati funzionali vengono applicati quando i dati consistono in osservazioni misurate con errore di funzioni casuali che possono differire tra i soggetti presi in esame. Per studiare l'eterogeneità tra gli individui, consideriamo il seguente modello gerarchico:

$$\begin{gathered} y_i(t) = f_i(t) + \varepsilon_i(t), \;\; \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2) \\ f_i \sim P \end{gathered} \tag{3.7}$$

dove $y_i(t)$ è una misura con errore della funzione f_i per l'individuo i al tempo t, $i = 1, \dots, n$ (cioè non è possibile osservare direttamente f_i, ma si hanno a disposizione le misure con errore $y_i(t)$ per $t \in \boldsymbol{t}_i = \left(t_{i1}, \dots, t_{in_i}\right)'$), $\varepsilon_i(t)$ è un errore di misura e P è una distribuzione sullo spazio delle funzioni $\mathcal{T} \to \Re$.

Nell'analisi di dati funzionali è usuale semplificare la modellazione assumendo che le ignote funzioni cadano nel tratto lineare di un insieme di basi di funzioni specificato a priori. Per esempio, per il modello (3.7) si ha:

$$f_i(t) = \sum_{h=1}^{p} \beta_{ih} b_h(t), \;\; \forall t \in \mathcal{T}$$

dove $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})'$ sono coefficienti delle basi specifici per il soggetto i e $\boldsymbol{b} = \{b_h\}_{h=1}^{p}$è un insieme di basi di funzioni.

Per evitare la scelta di uno specifico insieme di basi di funzioni, Bigelow e Dunson (2005) hanno proposto una modifica per permettere che il numero di nodi e la loro posizione sia ignota, ponendo una *a priori* DP sulla distribuzione dei coefficienti delle basi. In alternativa è possibile evitare del tutto il ricorso alla

specificazione mediante basi di funzioni delle funzioni casuali, seguendo l'approccio utilizzato, ad esempio, da Dunson e Herring (2006) e da Scarpa e Dunson (2009): si assume una mistura di processi di Dirichlet funzionali (FDP) per caratterizzare a priori $f_i(t)$ in cui la distribuzione di base del DP è un Processo Gaussiano.

Assumere una *a priori* FDP che ha come distribuzione di base un Processo Gaussiano (GP) significa assumere la seguente specificazione a priori per P: $P \sim DP(\alpha P_0)$, dove P_0 corrisponde a un GP. Quindi, nella rappresentazione *stick-breaking* in (3.2) specifichiamo $\theta_h \sim GP(\mu, C)$ e otteniamo che $(\theta_1, \theta_2, \dots, \theta_h, \dots)$ sono atomi funzionali, cioè funzioni casuali generate da un processo Gaussiano con media μ e funzione di covarianza C. Fissato un istante temporale t si ha che $f_i(t) \sim P(t)$ con $P(t) \sim DP(\alpha P_0(t))$, dove $P_0(t)$ è una distribuzione Gaussiana univariata.

Scegliendo come *a priori* un FDP, gli individui vengono raggruppati in cluster funzionali: indicando con $S_i = k$ l'appartenenza dell'individuo i al cluster k, si ha che $f_{S_i} = \theta^*_{S_i}$ per $i = 1, \dots, n$ dove θ^*_j è un'estrazione casuale dalla distribuzione di base che caratterizza la funzione specifica per il cluster j. Questo evita la specificazione di un set di basi di funzioni, tuttavia è necessario specificare le funzioni media e covarianza nel GP. È importante sottolineare che il numero di cluster funzionali e le stime *cluster-specific* dei parametri possono essere molto sensibili alla particolare scelta compiuta. Nell'implementazione del calcolo della distribuzione a posteriori, ovviamente non è possibile stimare la funzione su infiniti punti in $\mathcal{T}$. Nella pratica si procede alla stima su un insieme finito di punti che comprende gli istanti temporali in cui sono raccolti i dati. La distribuzione di base GP implica una misura di base normale multivariata sull'insieme finito di punti, perciò il calcolo della distribuzione a posteriori può procedere come quello per i modelli in cui si assume una distribuzione di base normale per il DP, con la particolarità che non vengono aggiornate media e covarianza (di dimensioni finite) nella distribuzione di base, ma si stimano i parametri che caratterizzano le funzioni media e covarianza. Per la funzione di covarianza solitamente è necessario un passo dell'algoritmo Metropolis-Hastings (Hastings, 1970), poiché la distribuzione a posteriori dei parametri non presenta forma chiusa.

4 Modello per l'analisi del traffico telefonico

Lo scopo di questo lavoro consiste nella previsione del *churn* a partire dai dati disponibili relativi alle caratteristiche dei clienti e al traffico telefonico. Data la particolarità dei dati emersa nelle analisi esplorative[1], desideriamo modellare congiuntamente l'andamento del traffico telefonico con l'indicatore di disattivazione.

Per raggiungere tale obiettivo è possibile estendere l'applicazione delle tecniche di analisi bayesiana non parametrica presentate nel Capitolo 3 alla modellazione congiunta di dati funzionali e variabili risposta. Seguiremo l'approccio proposto da Bigelow e Dunson (2009), apportando alcune modifiche: invece di specificare un modello gerarchico *multivariate adaptive spline* con una *a priori* DP per la distribuzione degli effetti casuali, assumiamo per la distribuzione delle traiettorie una *a priori* FDP con distribuzione di base GP. Inoltre, poiché il numero di cluster varia tra le iterazioni dall'algoritmo implementato per il calcolo della distribuzione a posteriori, sarà necessario annidare un algoritmo Metropolis-Hastings per l'aggiornamento dei parametri di gruppo la cui distribuzione a posteriori non presenta forma chiusa.

Il Paragrafo 4.1 presenta il modello congiunto sviluppato per l'analisi, nel Paragrafo 4.2 deriviamo la distribuzione a posteriori e nel Paragrafo 4.3 illustriamo gli *step* dell'algoritmo implementato per la stima. Infine, nel Paragrafo 4.4 affrontiamo alcuni problemi che sorgono nell'interpretazione dei risultati e presentiamo la soluzione adottata tra quelle proposte in letteratura.

4.1. Il modello

I dati a disposizione sono costituiti da n coppie di osservazioni $\{\boldsymbol{y}_i, z_i\}$, $i = 1, \dots, n$, dove $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$ rappresenta le osservazioni mensili relative alla traiettoria raccolte per T mesi consecutivi e z_i è la variabile risposta indicante

[1] Si veda il Capitolo 2

lo stato del cliente nel tredicesimo mese. Inizialmente specifichiamo un modello per la traiettoria, successivamente un modello per la risposta e infine integriamo le due componenti in un modello congiunto.

4.1.1. Modello per le traiettorie

Assumiamo per le osservazioni $\boldsymbol{y}_i$ il modello gerarchico (3.7):

$$y_i(t) = f_i(t) + \varepsilon_i(t)\,, \qquad f_i \sim G, \qquad \varepsilon_i(t) \sim \mathcal{N}(0, \tau^{-1}) \tag{4.1}$$

dove $y_i(t)$ è una misura con errore della funzione f_i per l'individuo i al tempo t, $i = 1, \dots, n$, $t \in \boldsymbol{t} = (t_1, \dots, t_T)'$, $\varepsilon_i(t)$ è un errore di misura che segue una distribuzione Gaussiana e G è una distribuzione sullo spazio delle funzioni $\mathcal{T} \to \Re$. Assumiamo, inoltre, che f_i e ε_i siano indipendenti.

Le caratteristiche dei dati longitudinali a nostra disposizione (ridotto numero di osservazioni per la traiettoria effettuate negli stessi istanti temporali per ogni individuo) portano alla specificazione di una *a priori* FDP con distribuzione di base GP per la distribuzione delle traiettorie latenti. Sotto questa assunzione, gli individui vengono automaticamente raggruppati in un numero non specificato a priori di cluster (o classi latenti) e le curve specifiche per ciascun cluster sono trattate in modo non parametrico.

Tale specificazione è stata adottata da Dunson e Herring (2006) nell'analisi dell'influenza dei livelli (variabili nel tempo) di un agente disinfettante nell'acqua potabile su alcune caratteristiche della gravidanza, come l'età gestazionale al momento del parto e il peso alla nascita.

Anche Scarpa e Dunson (2009), nell'analisi di raggruppamento delle curve relative alla temperatura basale durante il ciclo mestruale, hanno utilizzato un GP per la distribuzione delle funzioni casuali nella componente di contaminazione non parametrica del modello.

Scegliere un GP come distribuzione di base nel FDP evita la necessità di selezionare la base di funzioni mediante un algoritmo di tipo *reversible jump MCMC* (Green, 1995), pertanto, nel caso preso in esame, questo approccio risulta vantaggioso in termini di velocità computazionale e di interpretazione dei risultati.

Assumiamo, quindi, che G segua un DP con parametro di concentrazione α e con distribuzione di base un Processo Gaussiano (GP); in questo modo risulta fissata una *a priori* FDP per la distribuzione G:

$$f_i \sim G = \sum_{h=1}^{\infty} p_h \, \delta_{\Theta_h} \qquad \Theta_h \sim GP(\mu, C_\kappa) \tag{4.2}$$

dove i pesi di probabilità $\{p_h\}_{h=1}^{\infty}$ sono espressi nella forma *stick-breaking* come in (3.2) e $\boldsymbol{\Theta} = \{\Theta_h\}_{h=1}^{\infty}$ sono atomi funzionali, cioè funzioni casuali generate da un processo Gaussiano con media μ e funzione di covarianza C. Notiamo che la formulazione (4.1) - (4.2) corrisponde a una mistura infinita di processi Gaussiani.

Considerando la sequenza discreta di tempi $\boldsymbol{t} = (t_1, \dots, t_T)'$, si ha che $f_i(\boldsymbol{t}_i) \sim \mathcal{N}_T(\mu, C)$, cioè il GP induce una distribuzione normale multipla sui valori osservati del GP stesso.

La funzione di covarianza C controlla la tipologia delle forme osservate. Come semplice scelta per C assumiamo una funzione di covarianza esponenziale, poiché permette un'ampia varietà di forme funzionali; la forma esponenziale al quadrato può favorire eccessivamente funzioni lisciate, mentre scelte più flessibili, come la classe di funzioni di Matérn, sono caratterizzate da un numero elevato di parametri. Per la funzione di covarianza del GP assumiamo quindi la seguente forma esponenziale:

$$C_\kappa(t, t') = \kappa_1^{-1} \exp\{-\kappa_2^{-1}|t - t'|\}$$

dove κ_1 e κ_2 sono parametri ignoti (per le formulazioni alternative della funzione di covarianza si veda Rasmussen e Williams, 2006).

4.1.2. Modello per la variabile risposta

La variabile risposta z è una variabile dicotomica, che codifichiamo con i valori zero e uno:

$$z_i = \begin{cases} 1 & \text{se l'}i\text{-mo utente si è disattivato nel tredicesimo mese} \\ 0 & \text{altrimenti} \end{cases}$$

Data la natura dicotomica della variabile risposta è opportuno utilizzare un modello lineare generalizzato (GLM) per la modellazione. In particolare, assumiamo per z una distribuzione bernoulliana il cui valore atteso può essere diverso per ogni individuo e indipendente tra gli individui stessi. Specifichiamo, quindi, il seguente modello di regressione logistica (Dey, 2000):

$$z_i \sim Bin(1, \pi_i)$$
$$\pi_i = \frac{e^{\xi_i}}{1 + e^{\xi_i}} \tag{4.3}$$
$$\xi_i = a_i + \boldsymbol{x}_i'\boldsymbol{\gamma}$$

dove $\boldsymbol{\gamma}(p \times 1)$ è il vettore degli ignoti coefficienti di regressione e $\boldsymbol{x}_i(p \times 1)$ è il vettore contenente le osservazioni relative alle p variabili esplicative per l'individuo i. Poiché i dati a disposizione non comprendono informazioni "statiche" sui clienti da includere come covariate nel modello, si ha che $\xi_i = a_i$. Possiamo specificare una distribuzione a priori per l'effetto casuale a_i: $a_i \sim P$, assumendo per P una distribuzione normale oppure è possibile seguire un approccio semiparametrico, assumendo per P una distribuzione a priori DP con la scelta usuale della distribuzione di base normale.

4.1.3. Modello congiunto

Per permettere che la distribuzione della risposta vari in funzione della forma del predittore funzionale, modelliamo congiuntamente y_i e z_i, attraverso la specificazione di una distribuzione a priori per $\phi_i = \{f_i\,, a_i\}$:

$$\phi_i \sim H = \sum_{h=1}^{\infty} p_h \delta_{\Psi_h} \quad \Psi_h = \{\Theta_h\,, a_h^*\} \sim H_0 \tag{4.4}$$

dove i pesi $\{p_h\}_{h=1}^{\infty}$ sono definiti come in (3.2) e Ψ_h contiene le funzioni Θ_h unitamente all'effetto casuale a_h^* .

La misura di probabilità H_0 è definita come segue: $H_0 = GP(\mu, C_\kappa) \otimes \mathcal{N}(0, \upsilon^{-1})$; sotto questa assunzione Ψ_h è generato da due estrazioni indipendenti: $\Theta_h \sim GP(\mu, C_\kappa)$ e $a_h^* \sim \mathcal{N}(0, \upsilon^{-1})$. La specificazione a priori per ϕ_i è, quindi, la seguente:

$$\phi_i = \begin{pmatrix} f_i \\ a_i \end{pmatrix} \sim_{iid} H, \qquad i = 1, \dots, n$$

$$H \sim DP(\alpha H_0)$$

$$H_0 \sim \mathcal{N}_{T+1}\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} C_\kappa & 0 \\ 0 & \upsilon^{-1} \end{pmatrix}\right)$$

Completiamo la specificazione Bayesiana fissando la distribuzione a priori per i restanti parametri del modello congiunto:

$$v \sim Gamma(a_v, b_v)$$

$$\tau \sim Gamma(a_\tau, b_\tau)$$

$$\kappa_1 \sim Gamma(a_{k_1}, b_{k_1})$$

$$\kappa_2 \sim Gamma\left(a_{k_2}, b_{k_2}\right)$$

Poiché $\boldsymbol{y}$ e $\boldsymbol{z}$ sono indipendenti, la verosimiglianza congiunta può essere fattorizzata come il prodotto delle verosimiglianze per le due componenti del modello:

$$\begin{aligned} L(\boldsymbol{y}|f,\tau) &\propto \tau^{\frac{nT}{2}} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - f_i)'(\boldsymbol{y}_i - f_i)\right) \\ L(\boldsymbol{z}|\boldsymbol{a},v) &\propto \prod_{i=1}^{n} \exp(z_i a_i - \log(1+\exp(a_i))) \end{aligned} \tag{4.5}$$

Indichiamo con $\psi = (\psi_1, \dots, \psi_k)$ i $k \leq n$ valori di $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_h, h = 1, \dots, \infty)$ presenti nel campione degli n individui, con $\psi_h = (\theta_h, \tilde{a}_h)$, per $h = 1, \dots, k$. Si consideri inoltre l'indicatore di allocazione $S_i = h$ se l' i-mo individuo appartiene al cluster h, cioè se $f_i = \theta_h$ e $a_i = \tilde{a}_h$. Il modello definito dalle espressioni (4.1) e (4.3) può quindi essere riscritto come:

$$(y_i|S_i = h) \sim \mathcal{N}(\theta_h, \tau^{-1}) \text{ e } (z_i|S_i = h) \sim Bin\left(1, \frac{e^{\tilde{a}_h}}{1+e^{\tilde{a}_h}}\right) \tag{4.6}$$

Quindi le traiettorie per i soggetti sono estratte da una distribuzione Gaussiana centrata sulla media della traiettoria specifica per classe, mentre la variabile risposta è estratta da una distribuzione Bernoulliana con valore atteso espresso in funzione del parametro specifico per la classe.

Attraverso questa specificazione si permette alla media della distribuzione della risposta di variare tra i cluster delle traiettorie. In particolare, facendo riferimento ai dati di telefonia, il risultato che si ottiene da questa modellazione consiste in un insieme di cluster, che raggruppano i clienti con andamento del traffico e indicatore di disattivazione simile e, per ciascun cluster, è possibile calcolare una stima della probabilità di *churn*.

È opportuno precisare che la procedura adottata non separa nettamente i clienti attivi da quelli disattivati, individuando all'interno di ciascun gruppo le tipologie di comportamento prevalente. Una simile suddivisione sarebbe poco utile in fase di previsione del *churn*, quando l'azienda deve allocare un cliente ad un gruppo esclusivamente sulla base del traffico telefonico. Il risultato della modellazione congiunta, invece, permette di individuare i comportamenti comuni associando a ciascuno di essi la stima del rischio di *churn*. In questo modo, ogni mese l'azienda può allocare i clienti nei cluster avendo a disposizione l'andamento del traffico telefonico nei mesi precedenti e, in base all'appartenenza ai cluster con probabilità di *churn* più elevata, attuare delle specifiche azioni di *retention* solo sui clienti a rischio abbandono, massimizzando l'efficacia delle azioni intraprese.

4.2. Distribuzione a posteriori

In questo Paragrafo ci proponiamo di specificare la distribuzione a posteriori dei parametri del modello congiunto sviluppato nel Paragrafo precedente.

Indichiamo con $\boldsymbol{\psi}(\boldsymbol{t}) = \big(\psi_1(\boldsymbol{t}), \dots, \psi_k(\boldsymbol{t})\big)$, i $k \leq n$ valori distinti di $\{\phi_i(\boldsymbol{t}),\ i = n$. Definiamo l'indicatore di allocazione dell'individuo i al cluster h come segue: $S_i = h$ se $\phi_i(\boldsymbol{t}) = \psi_h(\boldsymbol{t})$, per $h = 1, \dots, k$. Inoltre, sia $\boldsymbol{S} = (S_1, \dots, S_n)'$ il vettore che indica l'allocazione di tutte le unità del campione. Escludendo l'unità i, si hanno $k^{(i)} \leq k$ valori distinti di $\{\phi_j(\boldsymbol{t}),\ j \neq i\}$ indicati con $\boldsymbol{\psi}^{(i)}(\boldsymbol{t})$, inoltre si ha che $S_j^{(i)} = h$ se $\phi_j(\boldsymbol{t}) = \psi_h(\boldsymbol{t})$ per $h = 1, .., k^{(i)}$ e $\boldsymbol{S}^{(i)} = \big(S_j, j \neq i\big)'$.

Poiché è impossibile trattare il modello con un numero infinito di parametri, integriamo rispetto all'ignota distribuzione H , ottenendo la seguente distribuzione a priori per $\big(\phi_1(\boldsymbol{t}), \dots, \phi_n(\boldsymbol{t})\big)$[2]:

$$\big(\phi_1(\boldsymbol{t}), \dots, \phi_n(\boldsymbol{t})\big)|\alpha, \kappa_1, \kappa_2, \upsilon \sim PU(\alpha, H_0)$$

Dopo l'integrazione rispetto a H, gli elementi $\phi_i(\boldsymbol{t}),\ i = 1, \dots, n$, sono interscambiabili (hanno la stessa distribuzione marginale e *full conditional*), ma non sono indipendenti. La distribuzione *full conditional* a priori per il generico elemento $\phi_i(\boldsymbol{t})$ dati è $\boldsymbol{S}^{(i)}, k^{(i)}, \boldsymbol{\psi}^{(i)}(\boldsymbol{t})$ è la seguente:

[2] si veda il Paragrafo 3.2

$$\left(\phi_i(\boldsymbol{t})\,|\boldsymbol{y}_i, z_i, \boldsymbol{S}^{(i)}, k^{(i)}, \boldsymbol{\psi}^{(i)}(\boldsymbol{t}), \alpha\right) \sim q_{i0} H_{i,0(\boldsymbol{t})} + \sum_{j=1}^{k^{(i)}} q_{ij}\, \delta_{\psi_j^{(i)}(\boldsymbol{t})}$$

dove $H_{i,0(t)} = \frac{L_i(\boldsymbol{y}_i, z_i|\phi_i(\boldsymbol{t}))H(\phi_i(\boldsymbol{t}))}{\int L_i(\boldsymbol{y}_i, z_i|\phi_i(\boldsymbol{t}))H(\phi_i(\boldsymbol{t}))d\phi_i}$ è la distribuzione *full conditional* a posteriori congiunta di $\{f_i, a_i\}$ sotto la distribuzione di base $H_{0(\boldsymbol{t})}$ (corrispondente a $\mathcal{N}_{T+1}\left(\mathbf{0}, \mathbf{C}_{\kappa,\upsilon}\right)$ con $\mathbf{C}_{\kappa,\upsilon} = \begin{pmatrix} C_\kappa(\boldsymbol{t}) & \mathbf{0} \\ \mathbf{0}' & \upsilon^{-1} \end{pmatrix}$ matrice diagonale a blocchi), e i pesi della mistura sono espressi da:

$$q_{i,j} \propto \begin{cases} \alpha h_i(\boldsymbol{y}_i, z_i) & se\ j = 0 \\ n_j^{(i)} L_i(\boldsymbol{y}_i, z_i|\psi_j) & se\ j > 0 \end{cases} \tag{4.7}$$

$$h_i(\boldsymbol{y}_i, z_i) = \int L_i(\boldsymbol{y}_i, z_i|\psi_j) dG_0(f_i, a_i) \tag{4.8}$$

Facendo riferimento alla fattorizzazione della verosimiglianza in (4.5) si ha che:

$$h_i(\boldsymbol{y}_i, z_i) = \int L_i(\boldsymbol{y}_i|f_i) dG_{0f}(f_i) \int L_i(z_i|a_i) dG_{0a}(a_i)\, h_i(\boldsymbol{y}_i, z_i) = h_i(\boldsymbol{y}_i) h_i(z_i)$$

dove $G_{0f} = \mathcal{N}_T(\mu, C_\kappa)$ e $G_{0a} = \mathcal{N}(0, \upsilon^{-1})$

La porzione relativa alla traiettoria ha forma chiusa:

$$h_i(\boldsymbol{y}_i) = \int \mathcal{N}_T\left(\boldsymbol{y}_i; f_j(\boldsymbol{t}), \tau^{-1} I_T\right) \mathcal{N}_T(f; \mu, C_\kappa) df$$

e risolvendo l'integrale otteniamo la distribuzione marginale di $\boldsymbol{y}_i$ rispetto a f_i :

$$\mathcal{N}_T(\boldsymbol{y}_i; \mu, \tau^{-1} I_T + C_\kappa)$$

La porzione relativa alla risposta non ha forma chiusa, poiché l'*a priori* non è coniugata con la verosimiglianza:

$$h_i(z_i) = \int \exp(z_i a_i - \log(1 + \exp(a_i))) \left(\frac{\upsilon}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\upsilon}{2} a_i^2\right) da_i$$

Bigelow e Dunson (2009) utilizzano un'approssimazione normale per il calcolo di $h_i(z_i)$, noi utilizziamo un'approssimazione numerica per il calcolo dell'integrale.

Gli effetti casuali $\{f_i, a_i\}$ possono essere estratti per ciascun soggetto, ma dal punto di vista computazionale è più efficiente sfruttare la proprietà di *clustering* di FDP, estraendo prima l'indicatore di cluster $\boldsymbol{S}$ e successivamente i distinti effetti casuali $\boldsymbol{\psi}$ dalla loro distribuzione *full conditional* (West, Müller e Escobar, 1994). L'indicatore di cluster per l'individuo i ha la seguente distribuzione *full conditional* a posteriori:

$$p\left(S_i = j \,|\boldsymbol{y}_i, z_i, \boldsymbol{S}^{(i)}, k^{(i)}, \psi^{(i)}(\boldsymbol{t})\right) = q_{ij} \quad \text{per } i = 1, \dots, n \tag{4.9}$$

dove i pesi di probabilità q_{ij} sono definiti nell'espressione (4.7). $S_i = 0$ implica la creazione di un nuovo cluster: l'individuo i viene assegnato al cluster $k = k^{(i)} + 1$, con $\phi_i(\boldsymbol{t}) = \psi_j(\boldsymbol{t}) \sim H_{i,0(\boldsymbol{t})}$.

All'estrazione degli indicatori di cluster segue l'aggiornamento di $\psi_j(\boldsymbol{t})$ dato $\boldsymbol{S}$. Per un fissato cluster j, $\psi_j = (\theta_j\ \tilde{a}_j)$. Per aggiornare ψ_j dalla distribuzione *full conditional*, generiamo casualmente un valore dalla sua distribuzione a priori aggiornata con i dati relativi agli individui del cluster j, per $j = 1, \dots, k$

$$p\big(\psi_j(\boldsymbol{t})|\boldsymbol{S}, \boldsymbol{y}, \boldsymbol{z}, \dots\big) \propto g_0\big(\psi_j\big) \prod_{i:S_i=j} L_i\big(\boldsymbol{y}_i, z_i|\psi_j(\boldsymbol{t}_i)\big)$$

Poiché le verosimiglianze per $\boldsymbol{y}$ e z sono indipendenti condizionatamente ai loro parametri, si ha:

$$p\big(\psi_j|\boldsymbol{S}, \boldsymbol{y}, \boldsymbol{z}, \dots\big) \propto \left(g_{0f}\big(\theta_j\big) \prod_{i:S_i=j} L_i\big(\boldsymbol{y}_i|\theta_j\big)\right)\left(g_{0a}\big(\tilde{a}_j\big) \prod_{i:S_i=j} L_i\big(\boldsymbol{z}_i|\tilde{a}_j\big)\right)$$

dove g_{0f} e g_{0a} sono le densità corrispondenti rispettivamente alle distribuzioni $G_{0f} = \mathcal{N}_T(\mu, C_\kappa)$ e $G_{0a} = \mathcal{N}(0, \upsilon^{-1})$

Quindi per ogni cluster j è possibile estrarre indipendentemente θ_j e $\tilde{a}_j$. La distribuzione a priori di base normale è coniugata con la verosimiglianza per i dati della traiettoria, quindi si ottiene la seguente distribuzione *full conditional* per θ_j:

$$p\big(\theta_j(\boldsymbol{t})|\boldsymbol{S}, \boldsymbol{y}, k\big) = \mathcal{N}\left(\left\{(\tau I) + m_j C_\kappa^{-1}\right\}^{-1}\left\{\tau I\mu + m_j C_\kappa^{-1}\bar{y}\right\}, \left\{(\tau I) + m_j C_\kappa^{-1}\right\}^{-1}\right) \tag{4.10}$$

La distribuzione *full conditional* per $\tilde{a}_j$ è la seguente:

$$p\big(\tilde{a}_j|\boldsymbol{S}, \boldsymbol{y}, \boldsymbol{z}, \dots\big) \propto \exp\left(-\frac{\upsilon}{2}\tilde{a}_j^2\right) \prod_{i:S_i=j} \exp\big(\, z_i\tilde{a}_j - \log\big(1 + \exp\big(\tilde{a}_j\big)\big)\big) \tag{4.11}$$

La distribuzione non presenta forma chiusa, pertanto è necessario l'algoritmo Metropolis-Hastings per estrarre valori da questa distribuzione.

4.3. Algoritmo

Per il calcolo della distribuzione a posteriori utilizziamo l'algoritmo *Gibbs sampler*, seguendo l'usuale schema per i modelli DP di mistura (MacEachern, 1998; MacEachern e Müller, 1998): si alterna l'aggiornamento dell'indicatore di allocazione nel cluster e il numero di cluster separatamente rispetto

all'aggiornamento dei parametri specifici per ogni gruppo. L'aggiornamento degli indicatori di cluster viene eseguito mediante l'algoritmo *Polya Urn Gibbs Sampling* (si veda Paragrafo 3.3).

Per l'estrazione dalle distribuzioni che non presentano forma chiusa, annidiamo un passo dell'algoritmo Metropolis-Hastings (Hastings, 1970). Per l'aggiornamento degli effetti casuali di gruppo, la cui distribuzione non presenta forma chiusa, non è possibile inserire esclusivamente uno *step* di tale algoritmo, ma è necessario annidare l'algoritmo completo all'interno del *Gibbs sampling*. Infatti, il numero di gruppi e la loro composizione varia tra le iterazioni *Gibbs sampling*, pertanto non possiamo utilizzare come valori iniziali per il Metropolis-Hastings le stime degli effetti casuali di gruppo ottenuti al passo precedente.

L'algoritmo è strutturato come segue[3]:

Step 0: *Inizializzazione*

Otteniamo i valori iniziali dei parametri mediante un'estrazione casuale dalla loro distribuzione a priori. Allochiamo la prima unità del campione al primo cluster.

Per $i = 1, \dots, n$ ripetiamo gli S*tep 1 e 2*:

Step 1: *Calcolo dei pesi q_{ij} per l'aggiornamento degli indicatori di cluster*

L'aggiornamento dell'indicatore di cluster per l'individuo i consiste nell'estrazione casuale di S_i dalla distribuzione *full conditional* (si veda l'espressione (4.9)), calcoliamo quindi i pesi di probabilità q_{ij} di tale distribuzione, espressi in (4.7):

per $j = 0$ otteniamo $h_i(\boldsymbol{y}_i)$ come estrazione casuale dalla distribuzione $\mathcal{N}_T(\boldsymbol{y}_i; \mu, \tau^{-1} I_T + C_\kappa)$, mentre $h_i(z_i)$ non ha forma chiusa, pertanto ricorriamo ad un'approssimazione numerica per il calcolo dell'integrale;

per $j > 0$ il calcolo dei pesi si riduce al calcolo della verosimiglianza per $\{\boldsymbol{y}_i, z_i\}$.

[3] Il codice R utilizzato per l'implementazione dell'algoritmo è riportato in Appendice A.

Step 2: *Aggiornamento dell'indicatore di cluster*

Aggiorniamo l'indicatore di allocazione, S_i, mediante un'estrazione casuale dalla distribuzione *full conditional*, sulla base del calcolo dei pesi q_{ij} effettuato nello *step* precedente:

$$p\left(S_i = j \,|\boldsymbol{y}_i, z_i, \boldsymbol{S}^{(i)}, k^{(i)}, \psi^{(i)}(\boldsymbol{t})\right) = q_{ij} \quad \text{per } j = 0, \dots, k^{(i)}$$

Se $S_i = 0$, l'individuo viene assegnato ad un nuovo cluster i cui parametri vengono estratti dalla distribuzione $H_{i,0(\boldsymbol{t})}$ (si veda Pagina 23).

Step 3: Aggiornamento della traiettoria media di gruppo

Per ciascuno dei cluster ottenuti nello *step* precedente, aggiorniamo i parametri relativi alla traiettoria di gruppo dalla distribuzione full conditional (4.10).

Step 4: *Aggiornamento degli effetti casuali di gruppo*

Poiché la distribuzione *full conditional* in (4.11) non presenta forma chiusa, è necessario annidare un algoritmo Metropolis-Hastings per simulare la distribuzione dalla quale estrarre casualmente il valore dell'effetto casuale per ciascuno dei cluster ottenuti nello *Step 2*.

Step 5: *Aggiornamento del parametro* τ

Aggiorniamo τ con un'estrazione dalla sua distribuzione *full conditional*:

$$p(\tau|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{S}) \sim Gamma\left(a_\tau + \frac{nT}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \theta_i)'(\boldsymbol{y}_i - \theta_i)\right)$$

Step 6: *Aggiornamento dei parametri* $\kappa_1 e\, \kappa_2$

La distribuzione *full conditional* per κ_1 e κ_2 è la seguente:

$$p(\kappa_1, \kappa_2|\theta_i, \dots) \propto |C_\kappa|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\theta_i - \mu)' {C_\kappa}^{-1}(\theta_i - \mu)\right) \times$$

$$\times {\kappa_1}^{a_{\kappa_1}-1} \exp\left(-b_{\kappa_1}\kappa_1\right) {\kappa_2}^{a_{\kappa_2}-1} \exp\left(-b_{\kappa_2}\kappa_2\right)$$

Poiché non presenta forma chiusa è necessario includere un passo di Metropolis-Hastings per l'aggiornamento dei parametri.

Step 7: Aggiornamento del parametro v

Aggiorniamo v con un'estrazione dalla sua distribuzione *full conditional*:

$$v|\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{S} \sim Gamma\left(a_v + \frac{k}{2}, b_v + \frac{\sum_{j=1}^{k} \tilde{a}_j^2}{2}\right)$$

Al termine della procedura iterativa otteniamo per ciascuna unità una stima della traiettoria del traffico telefonico e della probabilità di *churn*.

4.4. *Label switching* e *clustering*

Per raggiungere il nostro obiettivo, la previsione del *churn*, vorremmo utilizzare i risultati ottenuti per allocare le unità in gruppi, definiti in termini di traiettoria del traffico telefonico e probabilità di disattivazione. La principale difficoltà che sorge nell'interpretazione dei risultati dell'algoritmo MCMC è dovuta al fatto che il numero di cluster e la loro composizione varia tra le iterazioni dell'algoritmo. Questo problema, che in letteratura prende il nome di *label switching* (Redner e Walker, 1984), caratterizza i modelli di mistura in ambito bayesiano ed è causato dalla non identificabilità delle componenti sotto l'assunzione di distribuzioni a priori simmetriche. Sotto tale assunzione, infatti, la distribuzione a posteriori che ne risulta non varia rinominando (*relabelling*) le componenti della mistura. Questo comporta che le distribuzioni marginali a posteriori dei parametri siano identiche per ogni componente della mistura, perciò la stima delle quantità di interesse basata sull'utilizzo diretto dell'output MCMC risulta inappropriata.

Il *label-switching* non riguarda il calcolo di caratteristiche generali (come per esempio intervalli di credibilità e medie a posteriori) per la funzione $f_i(t)$ relativa al singolo individuo i, ma costituisce un serio problema nel caso in cui l'interesse principale dell'analisi sia il *clustering*.

Sono stati proposti diversi approcci per la soluzione del problema come, ad esempio, l'imposizione di vincoli di identificabilità (Diebolt e Robert, 1994) oppure i *relabelling algorithms* (Stephens, 1997a). Per una trattazione più ampia del problema del *label-switching* e delle possibili soluzioni si rimanda a Jasra, Holmes e Stephens (2005). In questa tesi seguiremo la proposta di Medvedovic e Sivaganesan (2002) per la formazione dei cluster: a partire dai risultati della

procedura iterativa costruiamo una matrice di distanza tra gli individui che costituisce la base per un metodo di raggruppamento gerarchico. Data la sequenza di *clustering* $(\boldsymbol{S}^1, \boldsymbol{S}^2, \dots, \boldsymbol{S}^N)$ generata dall'algoritmo, definiamo la distanza tra due individui i e i', cioè tra due coppie traiettoria-risposta, come la proporzione di campioni MCMC in cui i due individui sono allocati in cluster diversi:

$$D_{ii'} = \frac{\#\text{ campioni in cui } S_i \neq S_{i'}}{N}$$

Basandoci su questa distanza, costruiamo i gruppi di clienti simili per andamento del traffico telefonico e stato nel tredicesimo mese mediante una procedura di *clustering* gerarchico di tipo agglomerativo con legame completo (Everitt, 1993).

5 Applicazione: previsione del *churn*

5.1. Analisi dei risultati

Abbiamo applicato ai dati a disposizione il modello descritto nel Capitolo 4, scegliendo delle distribuzioni a priori poco informative poiché non abbiamo informazioni a priori sul comportamento dei clienti dell'azienda.

Assumiamo che la funzione media del processo gaussiano, μ, sia costante e pari al numero medio di telefonate effettuate dai clienti per ogni mese disponibile.

Dopo aver eseguito l'algoritmo per 5500 iterazioni, trascurando le prime 500 come periodo iniziale di *burn in*, abbiamo ottenuto i seguenti risultati.

I trace plot dei parametri riportati in Figura 5.1 non forniscono evidenze contro la convergenza della procedura iterativa.

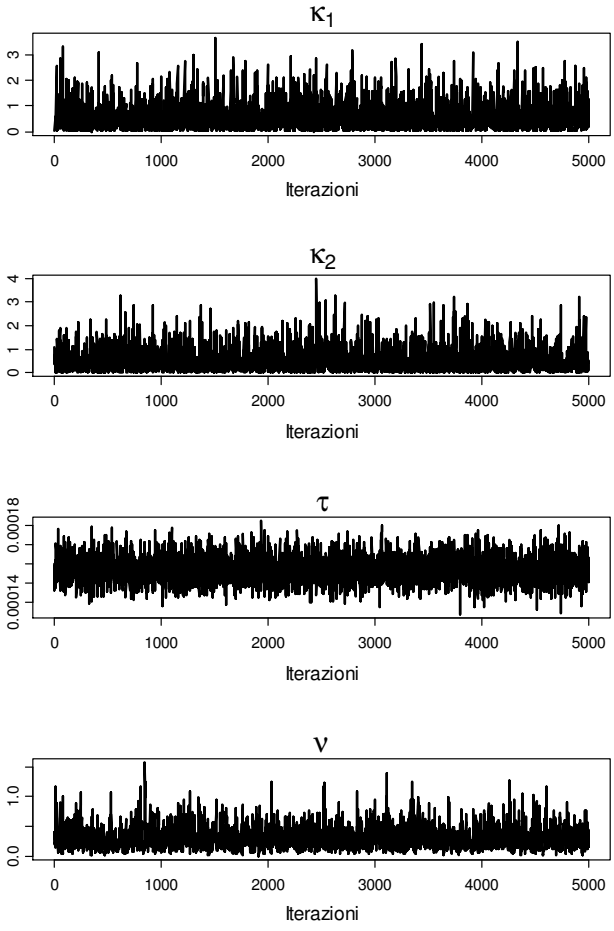


Figura 5.1 Trace plot parametri κ_1, κ_2, τ e υ

Riportiamo nella Tabella 5.1 alcune indicatori statistici che riassumono la distribuzione a posteriori dei parametri:

Parametro	Media	Deviazione standard	Primo quartile	Mediana	Terzo quartile
κ_1	0,507	0,498	0,143	0,352	0,719
κ_2	0,500	0,491	0,137	0,351	0,706
τ	0,0001572	0,0000062	0,0001532	0,0001570	0,000161
υ	0,292	0,172	0,170	0,261	0,381

Tabella 5.1 Sintesi della distribuzione a posteriori dei parametri κ_1, κ_2, τ e υ

Il risultato della procedura consiste in 5000 *pattern* di allocazione in gruppi per i clienti del campione e le relative stime dei parametri necessari per il modello.

Poiché il numero di gruppi individuati cambia tra le iterazioni dell'algoritmo di stima, in Figura 5.2 ne riassumiamo la distribuzione: presenta una certa simmetria e ha una media di circa 15 gruppi.

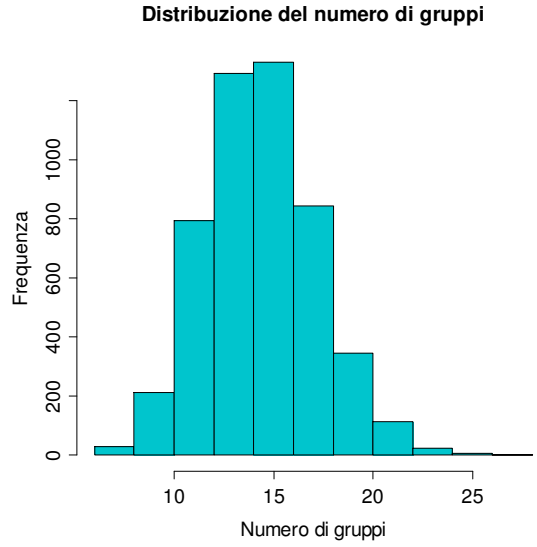


Figura 5.2 Distribuzione del numero di gruppi individuati in ogni iterazione

Non tutti i gruppi contengono un numero elevato di osservazioni, la Figura 5.3 mostra la distribuzione del numero di clienti per gruppo: il 25% dei gruppi è costituito da meno di dieci unità. Si nota, tuttavia, una concentrazione tra 1300 e 1800 unità per gruppo.

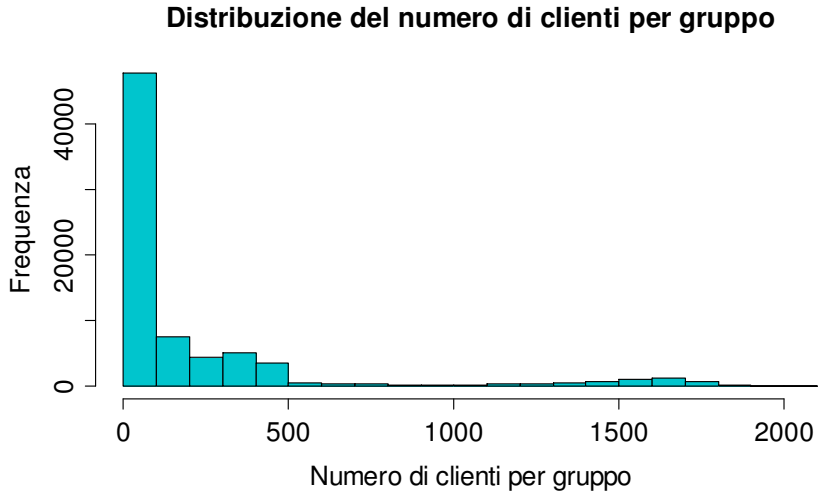


Figura 5.3 Distribuzione del numero di clienti allocati per ciascun gruppo

5.2. Clustering

Per la formazione dei cluster, seguiamo la procedura descritta nel Paragrafo 4.4. Inizialmente costruiamo la matrice di distanza tra coppie di clienti, utilizzando i risultati dell'algoritmo. In Figura 5.4 è riportata la distribuzione delle distanze tra coppie clienti: circa il 40% di tutte le distanze è prossimo a uno.

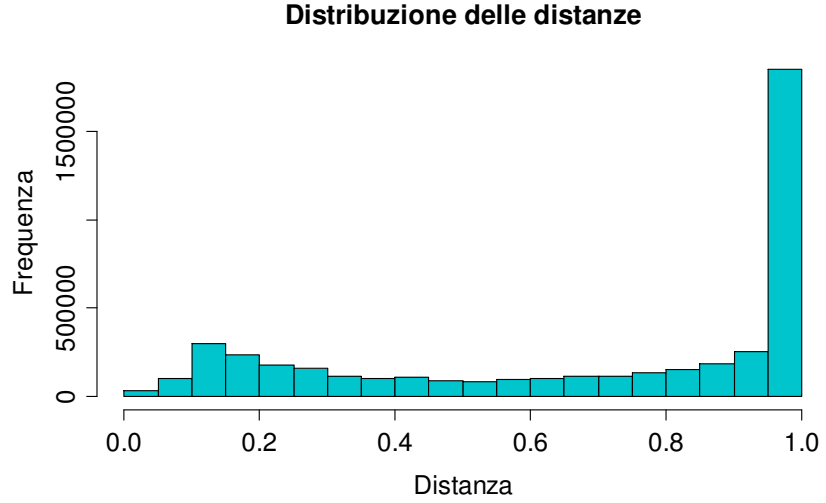


Figura 5.4 Distribuzione delle distanze tra coppie di clienti

Utilizziamo la matrice delle distanze come base per la procedura di *clustering* gerarchico con legame completo. La gerarchia di partizioni può essere rappresentata graficamente mediante una forma ad albero, chiamata

dendrogramma (si veda Figura 5.5), in cui gli individui sono rappresentati sull'asse orizzontale, mentre l'asse verticale indica la distanza tra i gruppi.

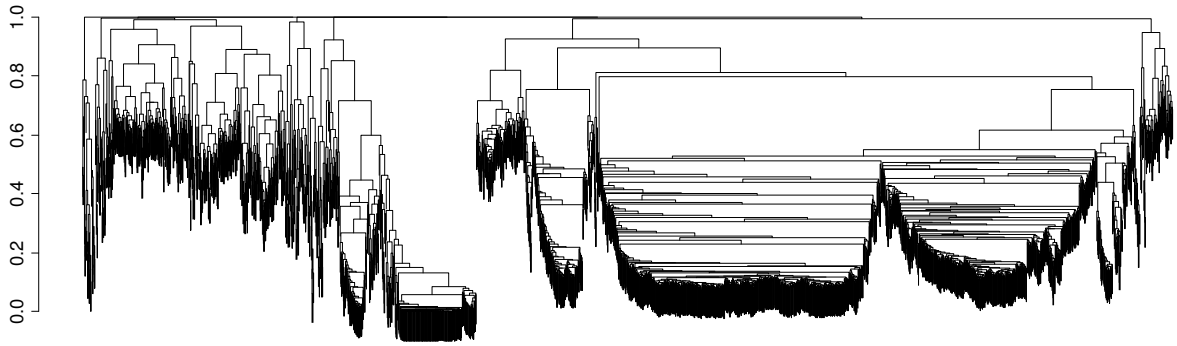


Figura 5.5 Dendrogramma

Per la determinazione del numero di gruppi, utilizziamo un criterio basato sulla differenza tra le misure di dissimilarità: sia $\delta_g = d_{g-1} - d_g$, dove d_g indica la distanza tra i due gruppi aggregati al passo della procedura che ha portato alla formazione di g gruppi. Il numero di gruppi ottimale coincide con quello per cui δ_g è massima, cioè quando si ha la differenza massima tra la distanza dei cluster aggregati in un passo del raggruppamento e la distanza tra i cluster aggregati al passo successivo[1]. Tale criterio ha portato all'individuazione di dodici gruppi.

La Figura 5.6 mostra la traiettoria media per ogni cluster calcolata come la media dei valori dei parametri ottenuti nelle 5000 iterazioni per gli individui appartenenti a ciascun cluster. Per alcuni gruppi si nota un andamento simile, ma su livelli diversi (ad esempio, gruppi 5 e 8). Inoltre, possiamo notare che in tutte le traiettorie è presente il "picco" negativo in corrispondenza dell'ottavo mese, già emerso nelle analisi esplorative, anche se con entità differenti tra i vari cluster. La flessibilità della modellazione attraverso un processo gaussiano ha permesso al modello di cogliere tale irregolarità.

[1] La procedura di cluster gerarchico adottata è di tipo agglomerativo, perciò procede aggregando successivamente gruppi precedentemente formati e con bassa dissimilarità tra di loro a partire da uno stato iniziale in cui ciascun individuo costituisce un gruppo a sé stante.

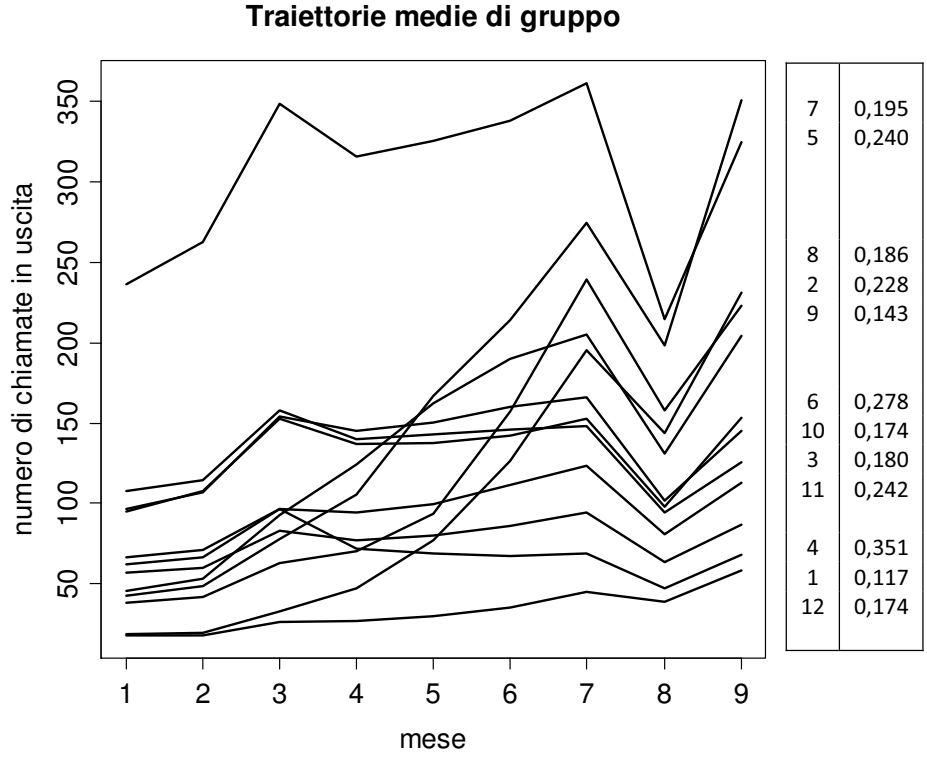


Figura 5.6 Traiettorie medie per ciascun cluster. La tabella a destra indica, per ciascuna traiettoria, il gruppo di riferimento e la corrispondente probabili di churn stimata secondo l'ordine dei valori dell'ultimo mese

Per ogni gruppo rappresentiamo in Figura 5.7 le traiettorie del traffico telefonico osservate, l'andamento medio stimato sulla base dei risultati del *clustering* e un intervallo di credibilità al 70% (linea a tratti). I gruppi sono ordinati in senso decrescente secondo la probabilità di *churn* stimata. Per agevolare il confronto tra i grafici dei vari cluster, abbiamo utilizzato la stessa scala per tutti i grafici.

Riportiamo, inoltre, nella Tabella 5.2 alcuni dati utili per il confronto tra i cluster individuati, quali la numerosità e la percentuale di clienti disattivati, la stima della probabilità di *churn* e l'intervallo di credibilità.

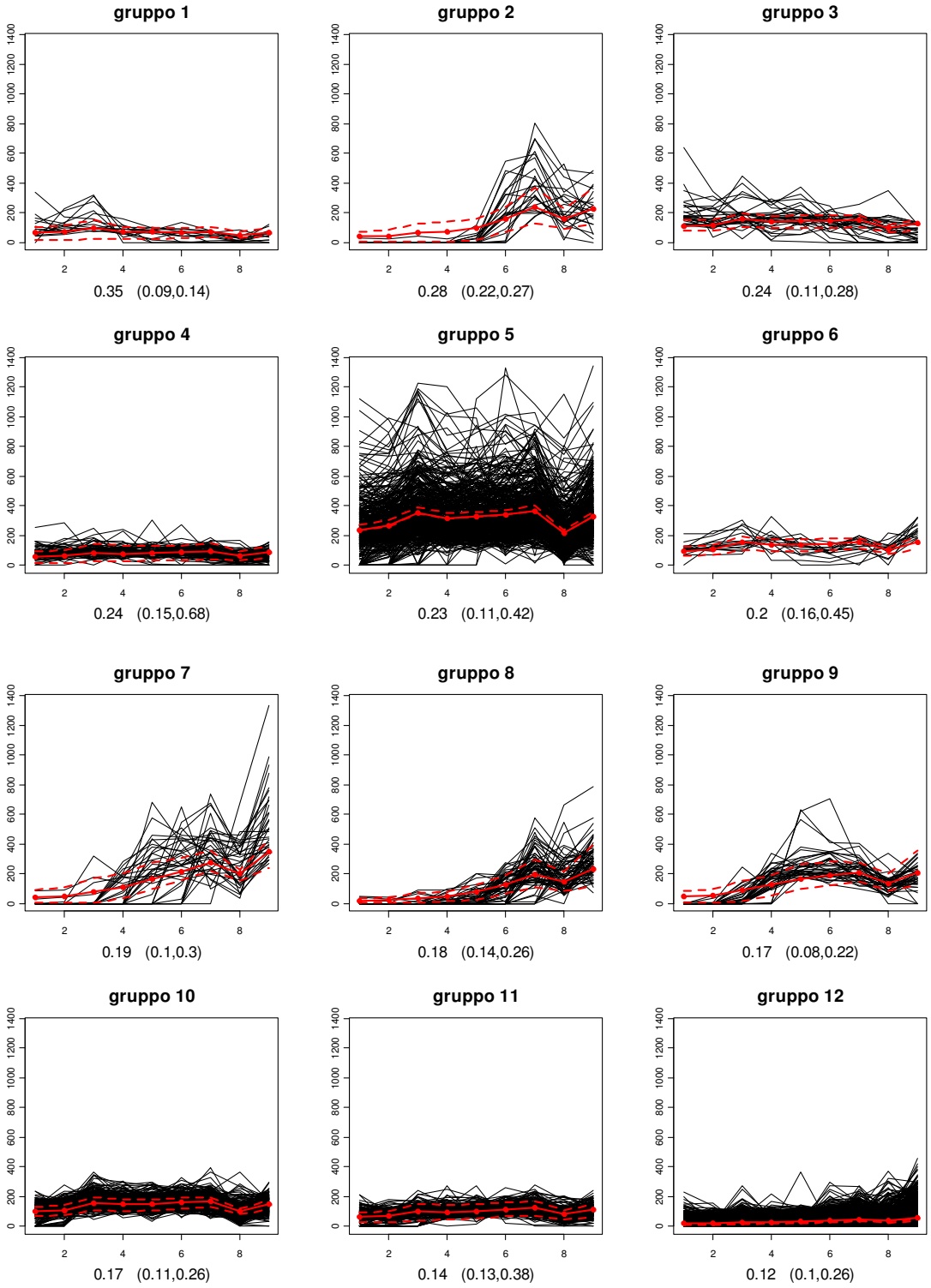


Figura 5.7 Rappresentazione delle traiettorie osservate per ogni cluster. Sotto ciascun grafico è riportata la probabilità di churn stimata dal modello e un intervallo di credibilità al 50%.

Cluster	Numerosità	Disattivati (%)	Probabilità di *churn* stimata	Intervallo di credibilità (50%)	
1	15	80,00	0,35	0,15	0,68
2	23	73,91	0,28	0,16	0,45
3	29	48,38	0,24	0,13	0,38
4	75	48,08	0,24	0,11	0,42
5	435	22,07	0,23	0,22	0,27
6	15	26,67	0,20	0,10	0,30
7	35	14,29	0,19	0,14	0,26
8	68	11,76	0,18	0,11	0,28
9	38	21,05	0,17	0,11	0,26
10	238	15,55	0,17	0,10	0,26
11	203	1,97	0,14	0,08	0,22
12	1826	11,06	0,12	0,09	0,14

Tabella 5.2 Confronto dei cluster

Dai risultati proposti emerge che i cluster che presentano probabilità di *churn* stimata più elevata sono quelli la cui percentuale di utenti realmente disattivati è maggiore (gruppi 1-4), nonostante il valore della stima non sia prossimo a uno (ciò è dovuto anche al fatto che il campione non è stato bilanciato).

Il gruppo 12 è il più numeroso, contiene circa il 61% delle unità del campione ed è caratterizzato da un andamento medio del numero di telefonate leggermente decrescente, il gruppo 7 presenta i valori medi di traffico più elevati.

In particolare, dall'analisi delle traiettorie che caratterizzano questi gruppi, notiamo che la procedura ha individuato andamenti diversi per chi abbandona l'azienda: nel caso del cluster 1, vediamo che gli utenti all'inizio del periodo di osservazione presentano un picco nel numero di telefonate effettuate, per poi calare. Mentre per il cluster 2 si ha un comportamento opposto: dopo i primi mesi in cui non vengono effettuate telefonate, nei mesi precedenti la disattivazione si ha un breve periodo di utilizzo maggiore, seguito da un calo. Entrambi questi comportamenti possono caratterizzare chi è a rischio abbandono: infatti, il primo può corrispondere ai clienti che riducono progressivamente il numero di chiamate fino a non effettuarne nessuna per alcuni mesi consecutivi, per poi abbandonare l'azienda. Il secondo andamento individuato potrebbe corrispondere a coloro che iniziano a utilizzare i servizi di telefonia alcuni mesi dopo l'attivazione, e

successivamente abbandonano l'azienda, probabilmente perché insoddisfatti del servizio offerto.

6 Estensione del modello

L'azienda possiede maggiori informazioni sui propri clienti rispetto a quelle utilizzate per le analisi nel Capitolo 5. Per esempio, per ciascun cliente si hanno più serie storiche relative al traffico telefonico: oltre al numero di chiamate in uscita, viene registrato il numero di chiamate in entrata, il numero di sms inviati, ecc. L'estensione del modello presentato nel Capitolo 4 per includere più traiettorie per ciascun cliente è immediata: è sufficiente replicare la stessa procedura per il numero di serie storiche disponibili.

L'azienda dispone, inoltre, delle informazioni "statiche" sui clienti, cioè quei dati raccolti non in forma di serie storica, come il tipo di contratto stipulato o le caratteristiche socio-demografiche del cliente. L'estensione del modello per l'aggiunta di questi dati non è diretta come nel caso precedente: è necessario apportare delle modifiche, per includere le variabili "statiche" come esplicative nel GLM per la risposta. In questo Capitolo descriveremo tali modifiche, sia per il modello sia per l'algoritmo implementato per il calcolo della distribuzione a posteriori e presenteremo i risultati ottenuti nell'applicazione del nuovo modello ai dati.

6.1. Modello congiunto

Consideriamo i dati costituiti da n terne di osservazioni $\{\boldsymbol{y}_i, \boldsymbol{x}_i, z_i\},\ i = 1, \dots, n$, dove $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$ rappresenta le osservazioni mensili relative alla traiettoria raccolte per T mesi consecutivi, $\boldsymbol{x}_i = \left(x_{i1}, \dots, x_{ip}\right)'$ rappresenta le osservazioni relative alle p variabili "statiche" e z_i è la variabile risposta indicante lo stato del cliente nel tredicesimo mese. Il modello per la traiettoria è specificato come segue:

$$\begin{aligned} y_i(t) &= f_i(t) + \varepsilon_i(t), \qquad \varepsilon_i(t) \sim \mathcal{N}(0, \tau^{-1}) \\ f_i &\sim G,\ G \sim DP(\alpha G_0), \qquad G_0 \sim GP(\mu, C_\kappa) \end{aligned} \tag{6.1}$$

dove $y_i(t)$ è una misura con errore della funzione f_i per l'individuo i al tempo t, $i = 1, \dots, n,\ t \in \boldsymbol{t} = (t_1, \dots, t_T)'$, $\varepsilon_i(t)$ è un errore di misura che segue una distribuzione Gaussiana e G è una distribuzione sullo spazio delle funzioni $\mathcal{T} \to \Re$.

Assumiamo che f_i e ε_i siano indipendenti.
Ora includiamo le variabili esplicative nel modello per la risposta e otteniamo il seguente modello di regressione logistica:

$$\begin{aligned} z_i &\sim Bin(1, \pi_i) \\ \pi_i &= \frac{e^{\xi_i}}{1 + e^{\xi_i}} \\ \xi_i &= a_i + \boldsymbol{x}_i'\boldsymbol{\gamma} \end{aligned} \tag{6.2}$$

dove $\boldsymbol{\gamma}(p \times 1)$ è il vettore dei coefficienti di regressione.

Per modellare congiuntamente la traiettoria con la risposta, assumiamo la seguente specificazione a priori per il modello:

$$\begin{aligned} &\phi_i = \begin{pmatrix} f_i \\ a_i \end{pmatrix} \sim_{iid} H, \qquad i = 1, \dots, n \\ &H \sim DP(\alpha H_0) \\ &H_0 \sim \mathcal{N}_{T+1}\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} C_\kappa & 0 \\ 0 & \upsilon^{-1} \end{pmatrix}\right) \\ &\tau \sim Gamma(a_\tau, b_\tau) \\ &\kappa_1 \sim Gamma(a_{k_1}, b_{k_1}) \\ &\kappa_2 \sim Gamma(a_{k_2}, b_{k_2}) \\ &\upsilon \sim Gamma(a_\upsilon, b_\upsilon) \end{aligned} \tag{6.4}$$

Sotto l'assunzione di indipendenza condizionata di $\boldsymbol{y}$ e $\boldsymbol{z}$, la verosimiglianza può essere fattorizzata nel prodotto delle verosimiglianze per le due componenti del modello:

$$\begin{aligned} L(\boldsymbol{y}|f, \tau) &\propto \tau^{\frac{nT}{2}} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - f_i)'(\boldsymbol{y}_i - f_i)\right) \\ L(\boldsymbol{z}|\boldsymbol{a}, \boldsymbol{X}, \boldsymbol{\gamma}, \upsilon) &\propto \prod_{i=1}^{n} \exp(z_i(a_i + \boldsymbol{x}_i'\boldsymbol{\gamma}) - \log(1 + \exp(a_i + \boldsymbol{x}_i'\boldsymbol{\gamma}))) \end{aligned} \tag{6.3}$$

Completiamo la specificazione Bayesiana fissando la distribuzione a priori per i parametri del GLM per la risposta:

$$\begin{aligned} &\boldsymbol{\gamma} \sim \mathcal{N}_p(\boldsymbol{\gamma}_0, diag(\boldsymbol{\eta})^{-1}) \qquad \boldsymbol{\eta} = (\eta_1, \dots, \eta_p)' \\ &\boldsymbol{\gamma}_0 \sim \mathcal{N}_p(0, (\omega)^{-1} I_p) \end{aligned}$$

$$\omega \sim Gamma(a_\omega, b_\omega)$$

$$\eta_l \sim Gamma\left(a_\eta, b_\eta\right) \quad l = 1, \dots, p$$

Assumiamo fissati tutti gli iperparametri delle distribuzioni a priori.

6.2. Algoritmo

Per il calcolo della distribuzione a posteriori dei parametri del modello specificato nel Paragrafo precedente, modifichiamo l'algoritmo presentato nel Paragrafo 4.3: la verosimiglianza in (4.5) viene sostituita con l'espressione (6.3) e, al termine della procedura, è necessario includere l'aggiornamento dei parametri relativi alle covariate del GLM. Ad eccezione dei coefficienti di regressione, $\boldsymbol{\gamma}$, le distribuzioni a priori degli altri parametri sono coniugate con la verosimiglianza, quindi si derivano agevolmente le distribuzioni a posteriori coniugate. Per l'aggiornamento dei parametri $\boldsymbol{\gamma}$ è necessario inserire uno *step* dell'algoritmo Metropolis-Hastings per l'estrazione dalla distribuzione a posteriori.

Lo *Step* 7 risulta così modificato:

Step 7: Aggiornamento dei parametri relativi alla risposta

Step 7a. Aggiornamento del parametro υ

Aggiorniamo υ con un'estrazione dalla sua distribuzione full conditional:

$$\upsilon|\mathrm{y}, \mathrm{a}, \mathrm{S} \sim \mathrm{Gamma}\left(\mathrm{a}_\upsilon + \frac{\mathrm{k}}{2}, \mathrm{b}_\upsilon + \frac{\sum_{\mathrm{j}=1}^{\mathrm{k}} \tilde{\mathrm{a}}_\mathrm{j}^2}{2}\right)$$

Step 7b. Aggiornamento del vettore di parametri $\boldsymbol{\gamma}$

Aggiorniamo i coefficienti di regressione con un'estrazione casuale dalla distribuzione *full conditional*:

$$p(\boldsymbol{\gamma}| \dots) \propto$$

$$\exp\left\{\sum_{i=1}^{n} z_i \boldsymbol{x'}_i \boldsymbol{\gamma} - \sum_{i=1}^{n} \log(1 + \exp(\boldsymbol{x'}_i \boldsymbol{\gamma} + a_i)) - \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)' diag(\boldsymbol{\eta})(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\right\}$$

Poiché la distribuzione non presenta forma chiusa, l'estrazione avviene mediante un passo dell'algoritmo Metropolis-Hastings.

Step 7c. Aggiornamento del vettore di parametri $\boldsymbol{\gamma}_0$

Aggiorniamo i parametri del vettore $\boldsymbol{\gamma}_0$ mediante un'estrazione dalla distribuzione *full conditional*:

$$\boldsymbol{\gamma}_0 | \ldots \sim \mathcal{N}_p \left(\left(\omega I_p + diag(\boldsymbol{\eta}) \right)^{-1} diag(\boldsymbol{\eta}) \boldsymbol{\gamma}, \left(\omega I_p + diag(\boldsymbol{\eta}) \right)^{-1} \right)$$

Step 7d. Aggiornamento del parametro ω

Aggiorniamo ω con un'estrazione dalla sua distribuzione *full conditional*:

$$\omega | \ldots \sim Gamma \left(a_\omega + \frac{p}{2}, b_\omega + \frac{\boldsymbol{\gamma}_0' \boldsymbol{\gamma}_0}{2} \right)$$

Step 7e. Aggiornamento dei parametri η_l

Aggiorniamo i parametri η_l, $l = 1, \ldots, p$ con un'estrazione dalla distribuzione *full conditional*:

$$\eta_l | \ldots \sim Gamma \left(a_\eta + \frac{1}{2}, b_\eta + \frac{1}{2} (\gamma_l - \gamma_{0l})^2 \right), \qquad l = 1, \ldots, p$$

6.3. Applicazione ai dati e risultati

Dal database originario abbiamo estratto le variabili "statiche" per ciascuna unità del campione: abbiamo a disposizione informazioni riguardanti il piano tariffario, il metodo di pagamento del servizio (tutti i contratti sono di tipo post-pagato) e l'età del cliente. Le prime due variabili sono di tipo fattore e sono schermate, ossia ciascun livello è indicato con un numero e non è possibile risalire ai livelli originari. L'età è una variabile continua che assume valori tra 15 e 55 anni, per le analisi l'abbiamo trasformata in fattore, costruendo quattro classi di età: 15-24 anni, 25-34 anni, 35-45 anni e 45-55 anni. La Figura 6.1 mostra la distribuzione in percentuale delle variabili "statiche" nel gruppo di clienti attivi e in quello dei disattivati.

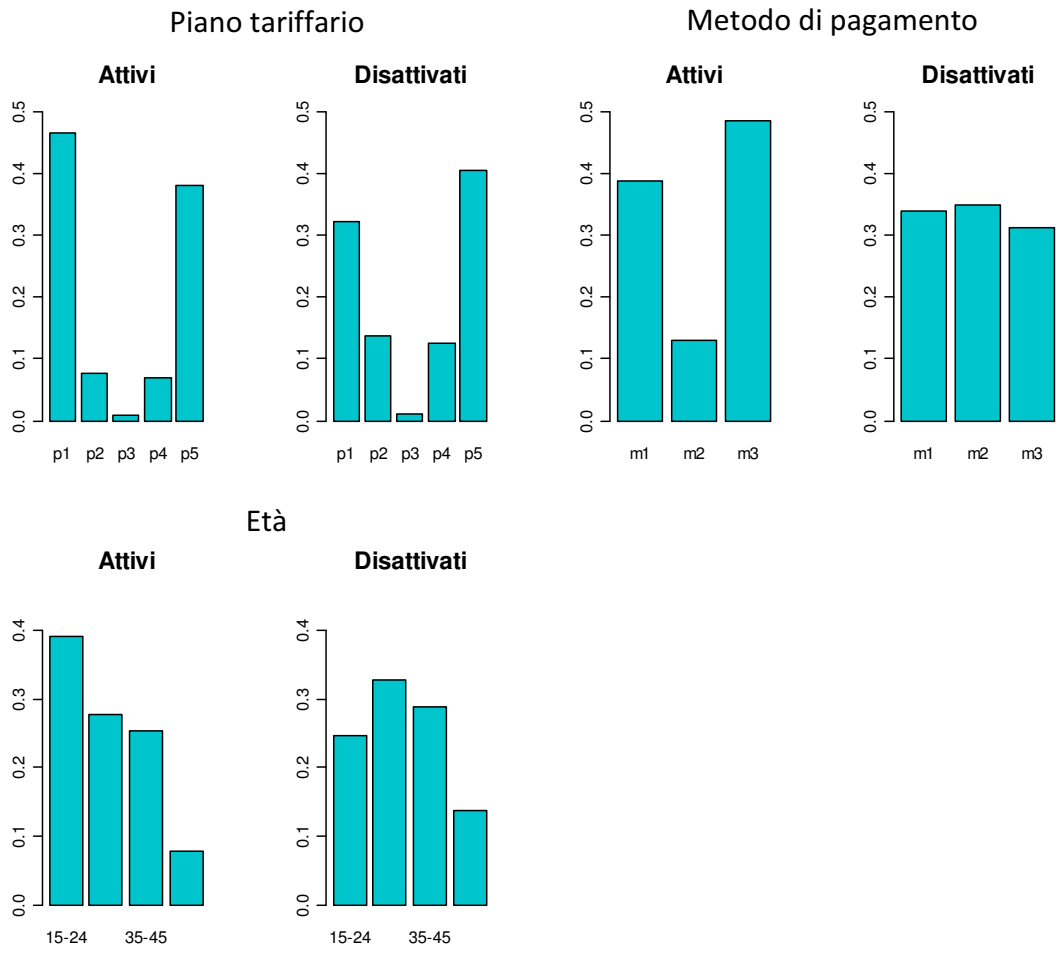


Figura 6.1 Distribuzione delle variabili "statiche" nei due gruppi (attivi e disattivati)

Per tutte le variabili si notano delle differenze nelle distribuzioni tra i due gruppi. In particolare per quanto riguarda il metodo di pagamento, i disattivi utilizzano i tre metodi disponibili in proporzioni simili, mentre gli attivi usano il metodo “m2” in misura inferiore rispetto gli altri. Inoltre, i clienti disattivati hanno una distribuzione dell’età centrata su valori più alti rispetto agli attivi.

Abbiamo eseguito l’algoritmo per 2700 iterazioni, escludendo le prime 700 come *burn in*. L’analisi dei *trace-plot* non ha fornito elementi contro la convergenza della procedura. Per includere nel modello le variabili fattore, abbiamo costruito delle variabili indicatrici, una per ciascun livello di ogni fattore e abbiamo escluso la prima per evitare problemi di multicollinearità. La Figura 6.2 mostra la distribuzione a posteriori dei parametri corrispondenti alle variabili *dummy* incluse nel modello.

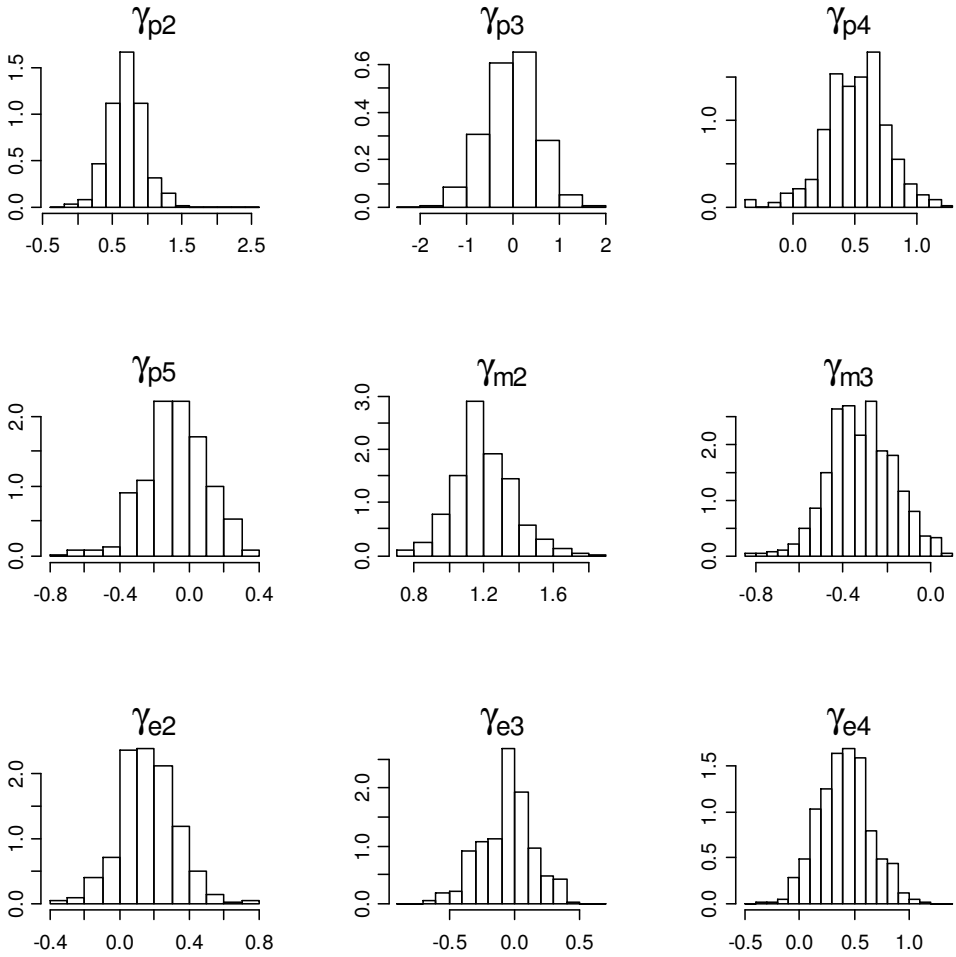


Figura 6.2 Distribuzione a posteriori dei coefficienti di regressione

La distribuzione del numero di gruppi individuati dalla procedura non presenta notevoli differenze rispetto alla precedente (si vedano Figura 6.3 e Figura 5.2).

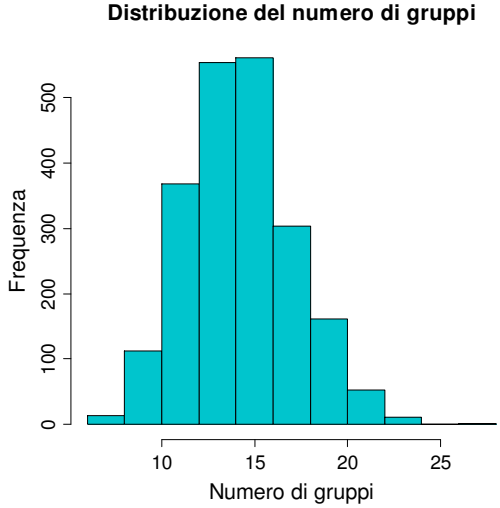


Figura 6.3 Distribuzione del numero di gruppi individuati dalla procedura

Abbiamo applicato ai risultati dell'algoritmo la procedura di clustering descritta nel Paragrafo 4.4, ottenendo il dendrogramma rappresentato in Figura 6.4.

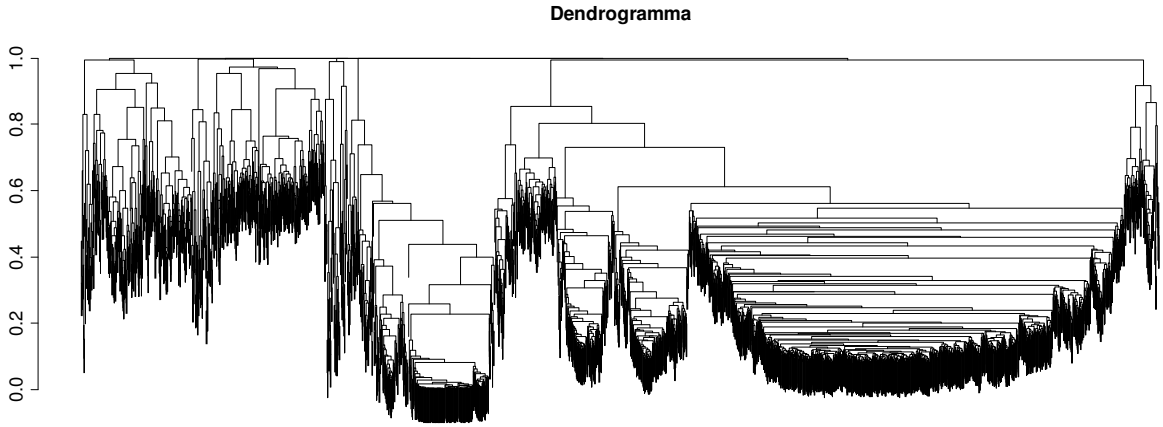


Figura 6.4 Dendrogramma

Per selezionare il numero ottimale di gruppi abbiamo utilizzato il criterio descritto nel Paragrafo 5.2, ottenendo sedici cluster. La Figura 6.6 mostra, per i nove cluster con probabilità stimata di churn più elevata, le traiettorie osservate, la traiettoria media e la probabilità di churn media di gruppo.

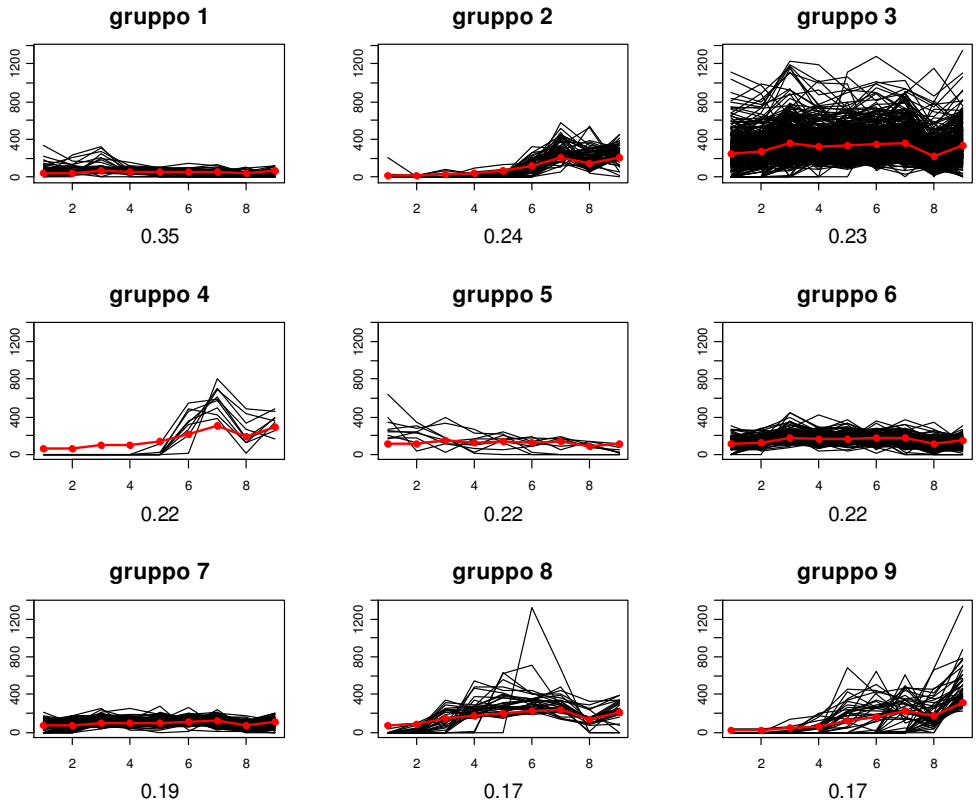


Figura 6.5 Rappresentazione dei cluster 1-9

La Figura 6.6 rappresenta i rimanenti sette cluster.

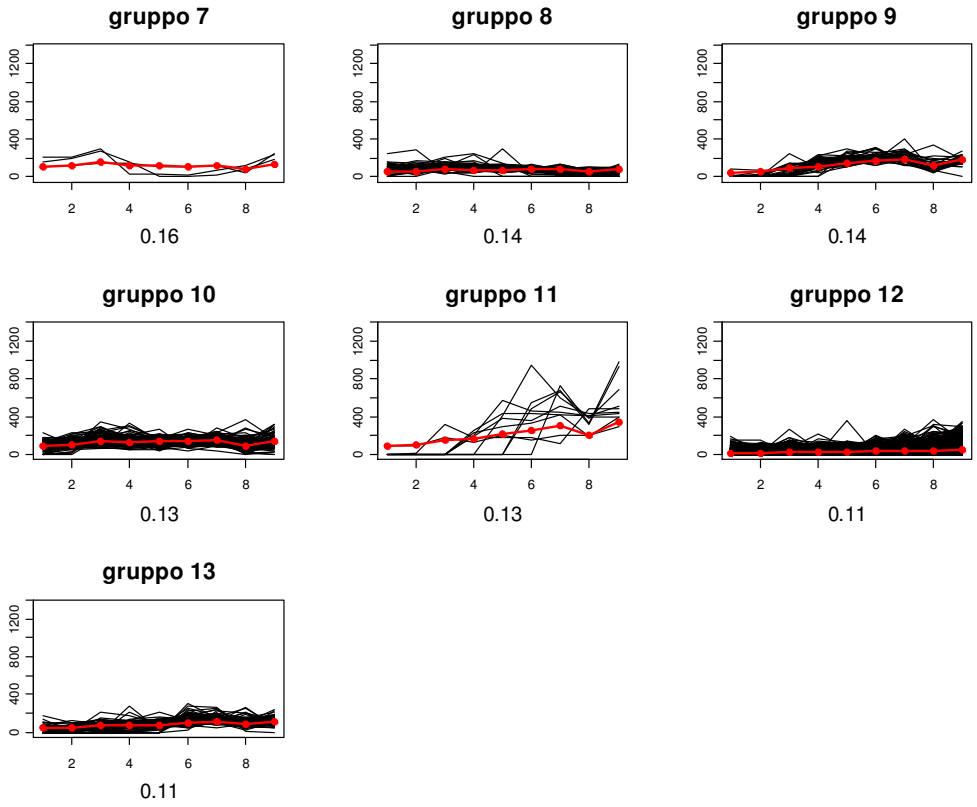


Figura 6.6 Rappresentazione dei cluster 10-16

Anche questo modello, come il precedente, riesce ad individuare i cluster con percentuale effettiva di churn maggiore (gruppi 1-4).

Confrontando i cluster con quelli ottenuti con il modello senza variabili esplicative nel GLM (Figura 5.7) notiamo che alcuni gruppi vengono individuati da entrambe le procedure: per esempio, i gruppi 1 e 3 sono corrispondono rispettivamente ai gruppi 1 e 5 di Figura 5.7. Questo significa che entrambi i modelli riescono a cogliere gli andamenti tipici del traffico telefonico e raggruppare i clienti secondo tali andamenti.

Poiché l'interesse principale dell'azienda è la previsione del churn, confrontiamo la capacità di previsione del nuovo modello con quella del modello stimato nel Capitolo 5. Per valutare la capacità di previsione del modello, uno strumento comunemente utilizzato è la funzione *lift* (Azzalini e Scarpa, 2004): essa fornisce una misura del miglioramento ottenuto dal modello nella previsione del *churn* rispetto alla classificazione casuale con probabilità uniforme pari alla proporzione di abbandono osservata nel campione.

La Figura 6.7 mostra le curve *lift* per i due modelli, calcolate sul campione di dati a disposizione.

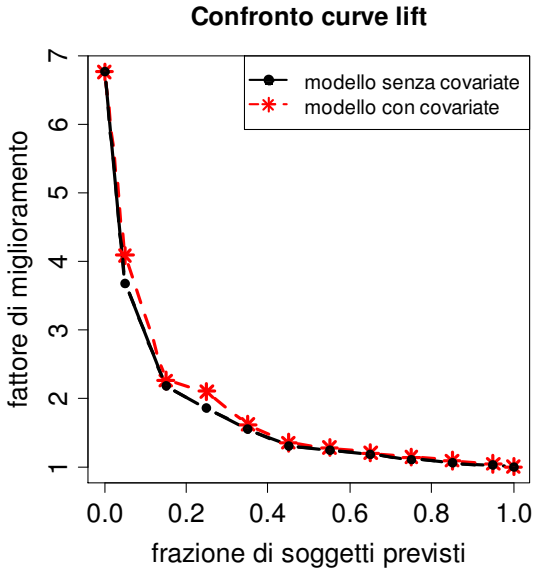


Figura 6.8 Confronto curve lift

Le due curve presentano un andamento molto simile, tuttavia, il modello che include le variabili esplicative nel GLM per la risposta mostra un fattore di miglioramento leggermente superiore: l'inserimento delle variabili "statiche" porta ad un miglioramento, seppur non elevato, della capacità di previsione del modello.

7 Conclusioni

Attraverso questa tesi abbiamo cercato di rispondere ad uno specifico problema di CRM che una società di telefonia incontra nella sua attività: la previsione del *churn*. Le caratteristiche dei dati a disposizione, quali il ridotto numero di mesi che compongono le serie storiche del traffico telefonico, la cadenza regolare e uguale per ogni cliente in cui sono state rilevate e l'assenza di specifiche informazioni a priori circa l'andamento del fenomeno oggetto di studio, ci hanno condotto alla scelta di un approccio bayesiano non parametrico. Abbiamo modellato il traffico telefonico come un dato funzionale, attraverso la specificazione per la sua distribuzione di una *a priori* Dirichlet *Process* scegliendo come base un Processo Gaussiano. Abbiamo successivamente incluso un modello GLM con effetto casuale per l'indicatore di disattivazione. Infine, abbiamo integrato le due componenti in un unico modello assumendo una distribuzione a priori congiunta per l'effetto casuale e il dato funzionale.

Il risultati ottenuti ci hanno permesso di individuare, nell'insieme del traffico telefonico degli utenti, gli andamenti comuni, a ciascuno dei quali abbiamo associato una stima della probabilità di *churn*. Abbiamo ottenuto due elementi utili per l'azienda: sia una classificazione degli utenti in base al loro profilo di utilizzo del servizio, sia un'indicazione utile per prevenire il *churn*. In questo modo, ogni mese l'azienda può allocare i clienti nei cluster avendo a disposizione l'andamento del traffico telefonico nei mesi precedenti e, in base all'appartenenza ai cluster con probabilità di *churn* più elevata, attuare delle specifiche azioni di *retention* solo sui clienti a rischio abbandono, massimizzando l'efficacia delle azioni intraprese.

L'estensione del modello, mediante l'inclusione di alcune variabili "statiche" (età del cliente, il piano tariffario e il metodo di pagamento adottato) come esplicative nel GLM per la risposta, ha permesso di migliorare la capacità di previsione del modello.

Appendice A: Codice R

Riportiamo il codice R utilizzato per l'implementazione dell'algoritmo descritto nel Paragrafo 4.3.

```
# numero di iterazioni
NSIM <- 5500

# alpha
alpha=1

dati <- as.matrix(read.table("dati.txt"))

library(mnormt)
source("funzioni_algoritmo.R")

t<-ncol(dati)-1
n<-nrow(dati)

# numero iterazioni per l'algoritmo Metropolis-Hastings annidato
nsim <- 1500

NUMGR<-rep(NA,NSIM)
A0 <- array(NA,dim=c(n,2,NSIM))
ACC_A <- array(NA,dim=c(NSIM,n/2))
PSI <- array(NA,dim=c(n/2,(t+1),NSIM))
TAU <- rep(NA,NSIM)
KAPPA <- array(NA,dim=c(NSIM,2))
ACC_K <- array(NA,dim=c(NSIM,2))
NI <- rep(NA,NSIM)
SS<-array(NA,dim=c(NSIM,n))
FI<-matrix(0,n,t)

# mu
mu<-colMeans(dati[,1:t]);mu

# Step 0: inizializzazione parametri

# kappa
a_k = c(1,1); a_k
b_k =c(1,1); b_k
kappa<-c(rgamma(1,a_k[1],b_k[1]),rgamma(1,a_k[2],b_k[2]));kappa

# Funzione di covarianza C
Cij<-matrix(0,t,t)
for (ci in 1:t){
     for (cj in 1:t){
     Cij[ci,cj] <- abs(ci-cj)
     }
```

```
}
C <- (1/kappa[1])*exp(-(1/kappa[2])*Cij)

# tau
a_tau <- 0.055
b_tau <- 1
set.seed (1516116);tau<-rgamma(1,a_tau,b_tau);tau

# ni
a_ni = b_ni =  0.5
ni <- rgamma(1,a_ni,b_ni); ni


# Ciclo per le iterazioni dell'algoritmo Gibbs sampler

for (w in 1:NSIM){

# STEP 1-2: calcolo pesi q_ij e aggiornamento degli indicatori
            di cluster

# inizializzazione

S <- numeric(n)
S[1] <- 1
idg <- 1
Q<-matrix(NA,n,n)
psi <- matrix(NA,nrow=(n/2),ncol=t+1)

psi[1,1:t] <- agg_theta0(i=1,dati,tau,C,mu)
A0[1,1:2,w] <- agg_a0.MH(i=1,nsim,dati,ni,a0=0)
psi[1,t+1] <- A0[1,1,w]

for (i in 2:n){
qi <-
c(fqij(i,idg,dati,S,psi,tau),fqi0(i,dati,tau,mu,C,ni,alpha,N=100
0))
Q[i,1:length(qi)] <- qi
S[i]<-sample(1:length(qi),1,T,prob=qi)
if (S[i]==length(qi)){
    psi[S[i],1:t] <- agg_theta0(i,dati,tau,C,mu)
    A0[S[i],1:2,w] <- agg_a0.MH(i,nsim,dati,ni,a0=0)
    psi[S[i],t+1] <- A0[S[i],1,w]
    idg<-1:length(qi)
   }
}

SS[w,]<-S


# Step 3: aggiornamento delle traiettorie medie di gruppo:

m<-table(S);m
k<-length(m)
NUMGR[w] <- k

psi[1:k,1:t] <- agg_theta(S,dati,tau,C,mu)
```

```
# Step 3: aggiornamento degli effetti casuali di gruppo:
aggiorn_a <-
sapply(1:k,agg_a.MH,nsim=nsim,dati=dati,S=S,ni=ni,a0=rep(0,k))
psi[1:k,t+1] <- aggiorn_a[1,]
ACC_A[w,1:k] <- aggiorn_a[2,]

PSI[,,w]<-psi


# Aggiornamento dei parametri relativi alla traiettoria

fi <- matrix(0,nrow=n, ncol=t)
for (i in 1:n)
fi[i,]<-psi[S[i],1:t]
FI<-((FI+fi))*(r>1)

# Step 5: aggiornamento tau
tau <- agg_tau(dati,fi,a_tau,b_tau)
TAU[w] <- tau


# Step 6: aggiornamento kappa e matrice di covarianza del GP
aggiorn_kappa <-
agg_kappa(a=rep(1,2),b=rep(1,2),kappa0=kappa,C,fi,mu,a_k,b_k)
kappa <- aggiorn_kappa$values
KAPPA[w,] <- kappa
ACC_K[w,] <- aggiorn_kappa$accepted


# aggiornamento della matrice di varianza e covarianza C
C <- (1/kappa[1])*exp(-(1/kappa[2])*Cij)


# Step 7: aggiornamento ni
ni <- agg_ni(a_ni,b_ni,a_tilde=psi[1:k,t+1])
NI[w] <- ni

}


# File "funzioni_algoritmo.R"


# Step 1 calcolo dei pesi q_i0
fqi0<-function(i,dati,tau,mu,C,ni,alpha,N)
{
  hi_y<-function(i)
  {
  require(mnormt)
  out<-dmnorm(dati[i,1:t],mu,(1/tau)*diag(t)+C)
  out
  }
    hi_z<-function(i,N)
    {
       d_hi_z <- function(a)
```

```
{dbinom(dati[i,t+1],1,prob=exp(a)/(1+exp(a)))*dnorm(a,0,sqrt(1/n
i))}
    cp <- rnorm(N,0,sqrt(1/ni))
    x=d_hi_z(cp)/dnorm(cp,0,sqrt(1/ni))
    out<-mean(x)
    out
    }
out <- alpha*hi_y(i)*hi_z(i,N)
out
}

# calcolo dei pesi q_ij per j>0
fqij <- function(i,idg,dati,S,psi,tau)
{
   L_i_psi <- function(i,idg)
   {
   p <- exp(psi[idg,t+1])/(1+exp(psi[idg,t+1]))
   require(mnormt)
   out<-dmnorm(dati[i,1:t],psi[idg,1:t],(1/tau)*diag(t))*
        dbinom(dati[i,t+1],1,prob=p)
   out
   }
   fm<-function(i,S,idg)
   {fs<-function(idg,S) sum(S[-i]==idg)
   sapply(idg,fs,S=S)
   }
   ff<-function(idg,i) (fm(i,S,idg)*L_i_psi(i,idg))
out <- sapply(idg,ff,i=i)
out
}


# Step 2: aggiornamento indicatori di cluster
# estrazione dei parametri per i nuovi cluster

agg_theta0<-function(i,dati,tau,C,mu){
invC <- solve(C)
varcov <- solve(tau*diag(t)+invC)
media <- varcov %*%(tau*diag(t)%*%mu + invC%*%(dati[i,1:t]))
out<-rmnorm(1,media,varcov)
out
}

agg_a0.MH<-function(i,nsim,dati,ni,a0)
{
   lposterior.j<-function(aj){
     out1 <- (-(ni/2)*aj^2) + aj*dati[i,t+1] - log(1+exp(aj))
     out1
   }
out<-numeric(nsim)
accepted<-0
a<-a0
for (u in 1:nsim)
{
       as<-a+rnorm(1)
```

```
       alpha=min(1, exp(lposterior.j(as)-lposterior.j(a)))
       uni <- runif(1)
       if (uni<alpha){
       accepted<-accepted+1
       a<-as}
       out[u]<- a
}
b <- 1:nsim/10
asim <- sample(out[-b],1)
out2<-c(asim,accepted/nsim)
out2
}

# Step 3: aggiornamento della traiettoria media di gruppo

agg_theta<-function(S,dati,tau,C,mu){
k <- length(m)
theta <- matrix(0,nrow=k,ncol=t)
invC <- solve(C)
  fv<-function(v) solve(tau*diag(t)+m[v]*invC)
  fm<-function(v){
      fv(v) %*% (tau*diag(t)%*%mu +
     m[v]*invC%*%colMeans(array(dati[S==v,1:t],dim=c(m[v],t))))}
 varcov<-sapply(1:k,fv)
 media<-sapply(1:k,fm)
fr <-function(b){
#set.seed(123)
require(mnormt)
rmnorm(1,media[,b],matrix(varcov[,b],t,t))
}
theta<-t(sapply(1:k,fr))
theta
}


# Step 4: aggiornamento degli effetti casuali di gruppo

# Metropolis-Hastings annidato

agg_a.MH<- function(j,nsim,dati,S,ni,a0)
  {
   lposterior.j<-function(aj){
     out1 <- (-(ni/2)*aj^2) + sum(aj*dati[S==j,t+1] -
log(1+exp(aj)))
     out1
   }
out<-numeric(nsim)
accepted<-0
a<-a0[j]
for (u in 1:nsim)
{
       as<-a+rnorm(1)
       alpha=min(1, exp(lposterior.j(as)-lposterior.j(a)))
       uni <- runif(1)
       if (uni<alpha){
       accepted<-accepted+1
```

```
       a<-as}
       out[u]<- a
}
b <- 1:nsim/10
asim <- sample(out[-b],1)
out2<-c(asim,accepted/nsim)
out2
}


# Step 5: aggiornamento tau
 agg_tau <- function(dati,fi,a_tau,b_tau)
{
out<-rgamma(1,a_tau+(n*t/2),b_tau+sum(apply((dati[,1:t]-
fi),1,crossprod)/2))
out
}


# Step 6:  aggiornamento kappa
agg_kappa<-function(a,b,kappa0,C,fi,mu,a_k,b_k)
{
lposterior <- function(kappa){
              fa<-function(a){
              out<-t(a)%*%solve(C)%*%a
              out}
              if (kappa[1]<0|kappa[2]<0) return(-Inf) else
              return(
              -n/2*log(det(C))-1/2*sum(apply((fi-mu),1,fa))+
              sum((a_k-1)*log(kappa)+a_k*log(b_k)-b_k*kappa))
}
out<-numeric(2)
accepted<-numeric(2)
kappa<-kappa0
   for (j in 1:2){
       kappas <- kappa
       kappas[j]<-rgamma(1,a[j],b[j])
       al=min(1, exp(lposterior(kappas)-lposterior(kappa)))
       u <- runif(1)
       if (u < al)
         {kappa[j]<-kappas[j]
         accepted[j]<-1}
       }
       out<- kappa
list(values=out,accepted=accepted)
}


# Step 7:  aggiornamento ni
agg_ni <- function(a_ni,b_ni,a_tilde)
{
rgamma(1,a_ni+(length(a_tilde)/2),b_ni+crossprod(a_tilde)/2)
}
```

Riferimenti Bibliografici

Antoniak, C. E. (1974). Mixtures of Dirichlet Process with applications to Bayesian nonparametric problems. *Ann. Statist. , 2*, 1152-1174.

Azzalini, A. e Scarpa, B. (2004). *Analisi dei dati e data mining.* Milano: Springer-Verlag.

Bigelow, J. L. e Dunson, D. B. (2009). Bayesian semiparametric joint models for functional predictor. *Journal of the American Statistical Association , 104*, 26-36.

Bigelow, J. e Dunson, D. B. (2005). *Semiparametric Classification in Hierarchical Functional Data Analysis.*

Blackwell, D. e MacQueen, J. (1973). Ferguson distributions via pólya urn schemes. *The Annales of Statistics , 1*, 353-355.

Bush, C. e MacEachern, S. (1996). A Semiparametric Bayesian model for randomised block designs. *Biometrika* , 275-285.

Dey, D. K. (2000). *Generalized linear models a bayesian perspective.* (D. K. Dey, S. K. Ghosh e B. K. Mallick, Eds.) New York: M. Dekker.

Diebolt, J. e Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B , 56*, 363-375.

Dunson, D. B. (2010). Nonparametric Bayes Application to Biostatistic. In N. L. Hjort, C. Holmes, P. Muller e S. G. Walker, *Bayesian Nonparametric.* New York: Cambridge University Press.

Dunson, D. B. e Herring, A. H. (2006). Semiparametric Bayesian trajectory models. *WP.06-16.* Durham, NC: Duke University, http://ftp.isds.duke.edu/WorkingPapers/06-16.pdf.

Dunson, D., Herring, A. e Siega-Riz, A. (2008). Bayesian inference on changes in response densities over predictor clusters. *Journal of the American Statistical Association , 103*, 1508-1517.

Escobar, M. D. e West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association, 90*, 577-588.

Everitt, B. S. (1993). *Cluster Analysis.* London: Edward Arnold.

Farinet, A. e Ploncher, E. (2002). Customer Relationship Management. Approcci e metodologie. Milano: Etas.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics , 1*, 209-230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics , 2*, 615-629.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika , 11*, 711-732.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika , 57*, 97–109.

Hjort, N. L. (2010). *Bayesian Nonparametrics.* (N. L. Hjort, C. Holmes, P. Muller e S. G. Walker, Eds.) Cambridge, UK;, New York: Cambridge University Press.

Ishwaran, H. e James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association , 101*, 179-194.

Ishwaran, H. e Zarepour, M. (2002). Dirichlet Prior Sieves in Finite Normal Mixtures. *Statistica Sinica , 12*, 941-963.

James, G. (2002). Generalized Linear Models With Functional Predictors. *Journal of the Royal Statistical Society. Series B , 64*, 411–432.

James, G. e Silverman, B. (2005). Functional Adaptive Model Estimation. *Journal of the American Statistical Association , 100*, 565–576.

Jasra, A., Holmes, C. C. e Stephens, D. (2005). Markov chain Monte Carlo methods. *Statistical Science , 20*, 50-67.

Lee, K. e Thompson, S. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine , 27*, 418-434.

MacEachern, S. (1998). Computational Methods for Mixture of Dirichlet Process Models. In D. Dey, P. Müller e D. Sinha, *Practical Nonparametric and Semiparametric Bayesian Statistics* (p. 23-44). New York: Springer-Verlag.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation , 23*, 727-741.

MacEachern, S. e Müller, P. (1994). Estimating normal means with a conjugate style Dirichlet process. *Communications in Statistics: Simulation and Computation , 23*, 727-741.

Medvedovic, M. e and Sivaganesan, S. (2002). Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles. *Bioinformatics , 18*, 1194–1206.

Muliere, P. e Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics , 26*, 283-297.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In T. S. Ferguson, L. S. Shapley e J. B. MacQueen, *Statistics, Probability and Game Theory* (p. 245-267). Institute of Matematical Statistics.

Ramsay, J. e Silverman, B. (1997). *Functional Data Analysis.* Springer.

Rasmussen, C. e Williams, C. I. (2006). *Gaussian Processes for Machine Learning.* MIT Press.

Redner, R. A. e Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review , 26* (2), 195–239.

Scarpa, B. e Dunson, D. B. (2009). Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors. *Biometrics , 65* (3), 772-780.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* (4), 639-650.

Stephens, M. (1997a). Bayesian Methods for Mixtures of Normal Distributions. *D.Phil. thesis* . Department of Statistics, University of Oxford.

West, M., Müller, P. e Escobar, M. (1994). Hierarchical priors and mixture models with application in regression and density estimation. In A. Smith e P. Freeman, *A Tribute to D. V. Lindley.* New York: Wiley.