Università degli studi di Padova

Facoltà di Statistica

Tesi di Laurea in

Statistica e informatica

Analisi bayesiana non parametrica del traffico telefonico

Relatore
Prof. Bruno Scarpa

Candidato
Angela Fraccalvieri

Anno Accademico 2009/2010

Università degli studi di Padova

Facoltà di Statistica

Tesi di Laurea in

Statistica e Informatica

Analisi bayesiana non parametrica del traffico telefonico

Relatore
Prof. Bruno Scarpa

Candidato
Angela Fraccalvieri

Anno Accademico 2009/2010

Indice

Capitolo 1

Introduzione

La segmentazione della *Customer Base* e in particolare il *profiling* dei comportamenti dei singoli clienti rispetto all'utilizzo e fruizione dei prodotti e servizi caratterizza la strategia di *marketing* di molte aziende. In particolare, le aziende caratterizzate da una continua interazione con i propri clienti (come banche, compagnie assicurative, società di telecomunicazioni, gestori di servizi internet ecc.) hanno implementato sofisticati strumenti di classificazione della clientela per migliorare la relazione con i singoli clienti, e per focalizzare al meglio le diverse strategie di marketing.
L'attenzione delle aziende si è focalizzata sul mercato e sul suo principale attore : il consumatore. Il consumatore si ritrova di fronte ad una vasta gamma di servizi e prodotti ed è solo una giusta conoscenza delle aspettative e delle necessità del cliente che può aiutare l'azienda ad acquisire sempre più utenti e a ridurre al minimo le quote di abbandono.
La stessa politica di marketing deve essere oggetto di un'attenta analisi per riuscire ad individuare le strategie più opportune e il giusto target di clientela interessato dalla singola strategia. È importante osservare che la percezione che il cliente ha dell'azienda si concretizza con le azioni di marketing intraprese e ad una errata politica corrisponde una valutazione negativa da parte dell'utente.
Questa crescente attenzione dell'azienda nei confronti del cliente può essere vista come un nuovo processo che coinvolge questo nuovo *network* relazionale composto da azienda e da clienti e che viene definito *Customer Relationship Management*, o più semplicemente *CRM*.
Gli strumenti maggiormente usati in questo contesto considerano le variabili disponibili principalmente in maniera statica, senza trarre vantaggio dalle caratteristiche dinamiche.
Queste ultime, invece, sarebbero di aiuto per descrivere il profilo dei diversi clienti, dal momento che il comportamento dei clienti è influenzato da fattori esterni che con il tempo si modificano. Grazie allo sviluppo informatico, la segmentazione della clientela, la classica strategia di marketing, ha potuto

avvalersi di algoritmi statistici sempre più complessi ed efficienti capaci di elaborare in poco tempo grandi quantità di dati. Grazie all'individuazione dei profili più ricorrenti tra i propri clienti è possibile rivolgersi ad essi massimizzando l'efficacia delle politiche di marketing intraprese.
La maggior sfida nell'applicazione delle tecniche statistiche per l'individuazione di cluster si ha nel settore telecomunicazioni in quanto queste aziende hanno a disposizione *datawarehouse* con una gran quantità di dati e informazioni relative a singoli clienti e ai singoli servizi forniti (Bisaglia, Gerolimetto, and Scarpa [2007]).
I modelli classici per l'analisi dei dati sono basati sull'analisi di tutti i dati a disposizione e utilizzando le usuali tecniche di *data mining*, non tengono conto della dipendenza temporale delle variabili.
L'obiettivo di questo elaborato è quello di riuscire ad individuare andamenti tipici, e di particolare interesse di marketing, nell'utilizzo del telefono cellulare, sfruttando l'informazione temporale dei dati longitudinali considerati. Sfruttando l'informazione temporale si riescono ad individuare i diversi modi che l'utente ha nell'utilizzare i servizi offerti.
In particolare, con l'approccio di analisi utilizzato, si vuole permettere una certa flessibilità nell'individuare le diverse tipologie di *pattern*. Si supponga di essere interessati ad individuare quegli utenti caratterizzati da un utilizzo del telefonino con una prima fase costante seguito da una fase intermedia di decrescita che infine porta ad una fase finale con un livello costante e inferiore a quello di partenza.
Gli istanti temporali di inizio e fine di queste fasi intermedie possono differire tra un utente e l'altro pur mantenendo invariato il profilo di utilizzo del telefonino. Tutti gli utenti con questo profilo apparterranno quindi ad un unico cluster.

1.1 Dati telefonia

I dati a disposizione si riferiscono al numero di telefonate inviate e al numero di *sms* e di *mms* inviati relativi alle singole *sim card* di una compagnia di telefonia mobile.
Le informazioni sono state raccolte con una granularità temporale mensile e l'orizzonte di osservazione inizia a Novembre 2004 e termina ad Aprile 2006. Questi stessi dati sono stati analizzati da Bonetto [2007] e Cassol [2009]. In particolare nel lavoro di Maela Bonetto, il cui scopo è quello di riuscire a modellare il *churn* sulla base dei dati statici e longitudinali, si ritrova un'analisi descrittiva dettagliata delle serie storiche relative al numero di telefonate effettuate e al numero di *sms* e *mms* inviati.
Nella analisi descrittive effettuate si riscontra che l'andamento medio nell'utilizzo dei queste tre tipologie di servizi assumono lo stesso andamento.

Nella tesi di Giacomo Cassol invece vengono modellate alcune variabili informative sul 'valore' che il cliente rappresenta per l'azienda, individuando delle classi di ammontare monetario del traffico telefonico e modellando l'appartenenza a ciascuna di queste classi con le usuali tecniche di classificazione supervisionata. Tra le tecniche che ha utilizzato si ritrovano gli alberi di regressione e l'analisi discriminante, ha inoltre applicato tecniche innovative quali la distanza autoregressiva (Corduas and Piccolo [2008]) che permette di valutare la distanza tra le serie storiche sulla base del processo autoregressivo ARIMA(p,q,d) sottostante i dati.
Dalle analisi descrittive condotte da Maela Bonetto, si riscontra che i diversi servizi di telefonia (numero di telefonate effettuate, numero di *ssm* e di *mms* inviati) presentano degli andamenti simile. In particolare si assiste ad un abbassamento nei livelli di utilizzo del cellulare nel mese di aprile con un successivo innalzamento. Nelle analisi che effettuerò nel presente elaborato sintetizzerò queste informazioni creando un unico dato longitudinale che riassume l'utilizzo dell'apparecchio telefonico.
Queste serie storiche rappresentano quindi l'input dell'analisi e la base per l'individuazione di appropriati dati funzionali.
Dopo un primo approccio al problema con le usuali tecniche statistiche per l'analisi delle serie storiche presenti in letteratura, sui dati a disposizione non si sono individuati dei pattern tipici nell'andamento del traffico telefonico, si è quindi deciso di condurre le analisi secondo la metodologia bayesiana semiparametrica. In questo modo si è potuto includere nel modello le informazioni a priori sull'andamento dei dati e sui parametri che definiscono questo andamento.
L'aspetto importante della metodologia utilizzata è l'integrazione nel modello bayesiano di due componenti distinte, una parametrica e una non parametrica, individuando quindi un modello miscuglio.
Il compito della componente parametrica è quello di riuscire a cogliere nei dati gli andamenti caratteristici, quelli più frequenti nella popolazione e di diretto interesso per eventuali azioni di marketing, mentre la componente non parametrica si occuperà di individuare gli altri possibili pattern marginali presenti nella popolazione.
Si è deciso in particolare di focalizzare l'analisi nell'individuazione di andamenti descrescenti con eventuali *plateaux* all'inizio e alla fine del periodo in cui il traffico decresce. Questa scelta è stata guidata da considerazioni di marketing, in quanto, spesso, l'attenzione delle compagnie telefoniche è rivolta all'individuazione di dei profili di clienti che possono far supporre, in base ad un calo osservato nell'utilizzo del telefono cellulare, un loro abbandono e adottare su di essi delle eventuali azioni di *retention*.

Capitolo 2

Classificazione di serie storiche

La classificazione di serie storiche è un ambito di ricerca ancora aperto. La problematica principale collegata a questi oggetti è la definizione di un'appropriata misura di distanza.
Spesso quando ci si trova di fronte all'analisi di serie storiche il primo obiettivo è quello di ridurne la dimensione, l'ampiezza, estraendo delle informazioni riassuntive come le componenti di trend e stagionalità oppure procedendo ad una rappresentazione lineare a tratti per rendere più efficiente l'algoritmo per calcolare la matrice delle distanze.
La ricerca di insiemi di serie storiche caratterizzate da pattern comuni è un importante problema statistico che si presenta in molti campi applicativi che vanno dall'economia e la finanza alle comunicazioni ed alla biomatematica.
L'analisi di similarità porta a riconoscere gruppi di serie storiche che hanno comportamenti simili. La similarità delle serie storiche segnala relazioni di causalità (oppure similarità strutturale) dei processi che le generano e permette di fare analisi più accurate.
La ricerca di similarità fra serie storiche è una pratica ricorrente della ricerca economica e finanziaria. Nel lavoro di Focardi [2001] si possono ritrovare tre esempi pratici di raggruppamento di serie storiche in ambito economico e finanziario:

- Il primo esempio riguarda la suddivisione del territorio degli Stati Uniti in regioni economiche effettuata dalla *Federal Reserve Bank of Philadelphia* basandola su criteri economici e statistici, scegliendo un insieme di variabili economiche che caratterizzano gli stati. Le variabili rappresentano quantità quali il reddito pro-capite, l'occupazione in settori non agricoli, il prodotto interno lordo dello stato. Di queste variabili è stato calcolato un indice rappresentativo per ciascuno Stato monitorato con cadenza mensile. La *Federal Reserve Bank of Philadelphia* ha applicato tecniche di raggruppamento a queste serie storiche

per verificare quali gruppi di stati abbiano comportamenti simili e possano formare una regione. Una volta individuati i gruppi di stati con caratteristiche simili si è valutata la contiguità territoriale degli stati appartenenti ai singoli gruppi e nella maggior parte dei casi si sono ritrovati gruppi ben definiti anche dal punto di vista territoriale, mentre altri stati sono stati eliminati dai gruppi di appartenenza a causa della lontananza territoriale. Le regioni trovate mostrano buoni livelli di omogeneità.

- Il secondo esempio riguarda applicazioni di misura del rischio di credito che coinvolgono un grande numero di contratti individuali. Al fine di misurare il rischio di credito è importante determinare quantità quali la probabilità individuale d'insolvenza e le correlazioni fra comportamenti creditizi. L'unità statistica in questo caso è rappresentata dai contratti individuali afferenti ai singoli istituti bancari. Sono state usate variabili economiche quali gli indici di bilancio e variabili legate al comportamento dell'azienda nei confronti della banca che eroga il credito. Potrebbero anche essere usate variabili qualitative ottenute codificando i rapporti d'agenzia.

- Un terzo esempio di analisi in cluster è costituito dall'analisi dei titoli azionari in settori, stili e temi d'investimento, con l'obiettivo di segmentazione del mercato azionario in settori. Tradizionalmente i mercati azionari sono stati segmentati in aree geografiche e settori merceologici, oppure usando indicatori quali la capitalizzazione o il tasso di crescita aziendale. A causa della globalizzazione e della formazione di grandi conglomerati, tali segmentazioni si sono rivelate, ad oggi, poco efficaci. Si è ricercato pertanto un nuovo modo di segmentare i mercati finanziari basato sull'analisi delle correlazioni empiriche fra titoli.

Vediamo perciò che il raggruppamento delle serie finanziarie è un processo di raggruppamento e discretizzazione delle correlazioni che soddisfa alcune esigenze fondamentali tra cui:

- identificare aggregati a cui corrispondano scelte politiche od economiche

- identificare aggregati a cui corrispondano particolari caratteristiche

- ridurre la dimensionalità dei problemi per renderli trattabili sotto il profilo matematico e statistico

- separare il rumore dall'informazione costruendo correlazioni stabili

2.1 Analisi longitudinale

I *dati longitudinali* compaiono in molte scienze applicate in cui spesso ci si confronta con insiemi di dati correlati nel tempo e nello spazio (Verbeke and Molenberghs [2000]). Si parla di dati longitudinali quando per una medesima unità statistica della popolazione di studio, si rilevano in diversi istanti temporali le manifestazioni di una o più variabili di interesse. Quando si parla di dati correlati, oltre ai dati longitudinali, ci si riferisce a diverse tipologie di dati strutturati come:

- *dati multivariati* che hanno ricevuto molta attenzione nella letteratura statistica e che vengono analizzati con tecniche statistiche multivariate quali la regressione multivariata e l'analisi della varianza multivariata. Un esempio di questo tipo di dati si ha quando per le singole unità statistiche vengono analizzate congiuntamente più caratteristiche.

- *dati raggruppati* che si ritrovano quando una popolazione statistica può essere divisa in famiglie e all'interno di ciascuna di queste famiglie per ciascuna unità statistica si rilevano determinate caratteristiche.

- *misure ripetute*[1] che si hanno quando una stessa caratteristica viene rilevata sotto differenti condizioni sperimentali

Per ciascuna tipologia di dati correlati si hanno delle tecniche statistiche più o meno idenee in relazione alle loro caratteristiche strutturali. Per esempio nei dati longitudinali a differenza di quelli multivariati, possiamo riscontrare la presenza di una ben definita struttura di varianza e covarianza, quindi in questo caso si potranno applicare dei *modelli lineari generalizzati*, in caso in cui la componente di covarianza non è ben strutturata si potrà approcciare l'analisi dei dati con modelli più flessibili come i *modelli lineari generalizzati misti*.
Una caratteristica fondamentale dei dati longitudinali che permette di approcciare la loro analisi da un punto di vista più dettagliato è la componente temporale. Grazie a questa dimensione è possibili modellare la variabilità del fenomeno non solo tra le unità che compongono la popolazione o tra le variabili oggetto di analisi, ma anche tra i diversi istanti temporali. In questo ambito si ritrova l'analisi delle serie storiche che come *focus* principale ha l'individuazione di strutture di auto-covarianza nelle serie storiche analizzate, cioè la variabilità del fenomeno considerando punti delle serie storiche ad istanti temporali ritardati (*referenza*).
In questi casi infatti abbiamo che la fonte di variabilità del fenomeno è dovuta a:

[1] Si nota che spesso in letteratura il termine osservazioni ripetute viene usato come sinonimo di dato longitudinale, mentre in questo caso viene fatta una distinzione tra i due tipi di dati.

- variazioni del fenomeno tra gli individui
- variazioni del fenomeno 'negli' individui, o meglio, correlazioni seriali

e son proprio queste due componenti di variabilità che si vogliono analizzare con l'analisi longitudinale.
Quando si lavora con dati longitudinali si ha per ciascuna osservazione un vettore Y di punti ordinati nel tempo, di solito per indicare la componente temporale del vettore si indicizza il vettore con la lettera t, $Y(t)$.
Data la componente temporale di questi studi spesso, specialmente negli studi epidemilogici, può accadere di dover lavorare con dati longitudinali non bilanciati nel senso che si hanno a disposizione un numero diverso di dati temporali per ciascuna osservazioni statistica oppure non si ha la stesso intervallo di tempo intercorrente da una osservazione all'altra delle stesso dato longitudinale.
In questi non si possono applicare le tecniche statistiche multivariate, poiché la base dati non è bilanciata temporalmente. Al posto di usare i dati longitudinali osservati si potrebbe procedere ad una loro analisi preliminare stimando per ciascun dato longitudinale un modello di regressione lineare, e utilizzare successivamente i parametri del modello lineare come input di successive analisi multivariate. Si parla in questo caso di $two-stage\ analysis$.
Spesso, infatti, si vuole studiare la relazione tra un dato longitudinale di riferimento (*outcome*), Y, e altri dati longitudinali considerati esplicativi del nostro *outcome*, X, e un classico approccio per questo tipo di analisi è costituito dai *modelli a effetti misti*, con cui si studiano come i coefficienti β degli effetti cambiamo tra le unità statistiche.

2.1.1 Modello lineare ad effetti misti

Si supponga il dato longitudinale di interesse, relativo ad una popolazione di n individui, sia rappresentato della variabile casuale $Y_i(t_i)$, dove $i = 1, .., n$ rappresenta la singola unità statistica e $t_i = 1, ..., T_i$ rappresenta l'istante temporale di osservazione, dove l'ultimo istante temporale T_i può variare per le diverse unità statistiche.
Indicando con $\mathbf{Y_i}$ il vettore del dato longitudinale relativo alla $i-esima$ osservazione, un modello lineare ad effetti misti è dato da:

$$\mathbf{Y_i} = X_i\beta + Z_i b_i + \varepsilon_i$$

dove:

$X_i =$ matrice (T_i x q) delle variabili esplicative, o covariate, note
$Z_i =$ matrice (T_i x p) delle variabili esplicative, o covariate, note
$\beta =$ vettore degli effetti fissi
$b_i =$ vettore degli effetti casuali soggetto specifici
$\varepsilon_i =$ residui del modello

Si assume che i residui del modello siano indipendenti e seguono una distribuzione normale multidimensionale di media zero e matrice di varianze e covarianze pari a $\sigma^2 I_{T_i}$, dove I_{T_i} è una matrice identità $T_i - dimensionale$.
Anche i vettori dei parametri b_i si assume che siano indipendenti e identicamente distribuiti secondo una variabile casuale normale $q - dimensionale$ di media zero e matrice di varianze e covarianze Ω.
Segue che condizionatamente al vettore degli effetti b_i, $\mathbf{Y_i}$ si distribuisce normalmente con media $X_i\beta + Z_i b_i$ e matrice di varianze e covarianze $\sigma^2 I_{T_i}$.
Con questo tipo di modelli si assume che il vettore $\mathbf{Y_i}$ per ciascuna unità statistica segue un modello di regressione lineare in cui alcuni parametri di regressione sono specifici della popolazioni e altri specifici della singola unità statistica.
Con questo approccio è possibile modellare le serie storiche in funzione di una covariata che esprime in tempo e capire in questo modo come evolve il fenomeno secondo un approccio parametrico e può quindi essere usata per capire quali sono i trend principali dell'insieme di dati longitudinali considerato.
Ma può succedere che all'interno di uno stesso fenomeno, ci siano alcune manifestazioni di dati longitudinali il cui andamento si discosta da questa forma parametrica e per cui i modelli lineari a effetti misti non possono essere sufficienti.
Per questo motivo nel Capitolo 3 vedremo come è possibile affrontare lo studio dei dati longitudinali sotto un'impostazione semiparaemtrica.

2.2 Riduzione delle dimensionalità

Notiamo che il problema di definire la similarità fra serie si trova anche nel contesto dei database e *data-warehouse*. Infatti è stato posto da tempo il problema di fare ricerche per similarità in una base dati di oggetti ad alta dimensionalità. Questo è un problema classico di *pattern recognition* che ha condotto a metodologie applicative nuove in domini quali la biomatematica, la sismologia, le telecomunicazioni ed i problemi di *intelligence* quali il riconoscimento di esplosioni nucleari a partire da *pattern* sismologici.
Tra le tecniche presenti in letteratura per esplorare le serie storiche e trovare andamenti simili si ritrovano per esempio il *Dynamic Warping Time* (DTW), la *Singular Value Decomposition* e le *Wavelet*.
Le ultime due tecniche si occupano in particolare di ridurre la dimensione della serie storica estraendo delle informazioni riassuntive delle serie storiche e utilizzando solo queste ultime per le successive analisi dei dati.
Il processo di *wavelet analysis* (Graps [1995]) sfrutta alcune proprietà delle serie storiche viste come funzione del tempo. L'analisi consiste nell'adottare una funzione prototipo *wavelet* chiamata *analyzing wavelet* o *mother wavelet*, e rappresentare la serie storica iniziale in termini di espansione di *wavelet*,

utilizzando una combinazione lineare di funzioni *wavelet*. Le operazioni sulle serie storiche possono essere quindi eseguite utilizzando i coefficienti corrispondenti a questa combinazione lineare. Se si filtrano i coefficienti al di sotto di un prefissato valore, si possono utilizzare le *wavelet* come una strumento per la riduzione della dimensione delle serie storiche.
Anche la *Singular Value Decomposition* (Korn, Jagadish, and Faloutsos [1997]) è una tecnica matematica per ridurre la dimensione dei vettori.
Il principio di base è l'individuazione del minor numero di *singular value* in grado di spiegare la variabilità dei dati.
Se indichiamo con X la matrice dei dati, possiamo ottenere la seguente rappresentazione:

$$X = UDV^t$$

Sulla diagonale principale della matrice D vi sono i *singular value* in ordine crescente, selezionando i *singular value* con valore maggiore otteniamo una sotto matrice D e di conseguenza consideriamo le prime righe delle matrici U e V in modo da ottenere una matrice X di dimensioni ridotte.
Il *Dynamic Warping Time* invece permette di confrontare direttamente l'intera serie storica con altre serie storiche per valutarne il grado di somiglianza.
Ai fini dell'individuazione di andamenti simili nel traffico telefonico e dato che le dimensionalità delle serie oggetto di studio non rappresenta un problema, si è provato ad applicare il *Dynamic Warping Time* ai dati sulla telefonia. Si vedrà nel prossimo paragrafo come lavora questo algoritmo.

2.3 Dynamic Time Warping

Per poter trovare degli andamenti tipici, cioè delle similarità, tra le serie storiche il primo passo dell'analisi è quello di definire una misura di distanza tra serie storiche.
La distanza più utilizzata negli algoritmi di classificazione è la distanza Euclidea, ma questa distanza, applicata alle serie storiche si rivela eccessivamente sensibile a piccoli scostamenti dei valori sull'asse del tempo. Quindi serie storiche che hanno visibilmente un andamento simile, ma che non sono perfettamente allineate sull'ascissa sono caratterizzate da una distanza Euclidea elevata.
Questo perché, come schematizzato in Figura 2.1 con la distanza Euclidea si valuta la distanza tra i punti con la medesima coordinata temporale.

Nello studio delle similarità tra serie storiche risulta infatti interessante individuare andamenti strutturalmente simili anche se manifestati in un intervallo temporale diverso.
Un'intuitiva soluzione a questo problema è quello di confrontare i punti di una serie A non solo con il punto corrispondente sull'asse del tempo della serie B, ma anche con i punti adiacenti, scegliendo come distanza il minimo

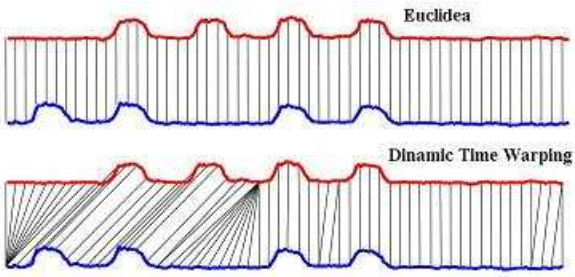

Fig. 2.1: Metodo di confronto di due serie storiche con la distanze Euclidea e con la distanza *Dynamic Time Warping*

tra queste distanze.
Il *Dynamic Time Warping* è un algoritmo che permette di valutare la distanza tra serie confrontando i valori di una serie con un sottoinsieme di altri valori delle altre serie.
Per una descrizione dettagliata di questo algoritmo si riporta all'articolo di Giorgino [2009] di cui si sfrutta l'algoritmo implementato in R per l'applicazione dell'algoritmo ai dati.
Questo algoritmo permette di calcolare la distanza *Dynamic Time Warping* associata al *warping path* più breve che si ottiene confrontando una coppia di serie storiche. Per avere una visione grafica di come lavora l'algoritmo, si supponga di creare una griglia come quella in Figura 2.2 che permette di mappare i possibili incroci tra i punti delle serie storiche A e B considerate. Il *warping path* si ottiene partendo dalla prima cella in basso a sinistra che confronta il primo punto della serie A e della serie B e individuando man mano sulla griglia le celle contigue tra loro che individuano i punti delle due serie caratterizzati da una distanza minore.

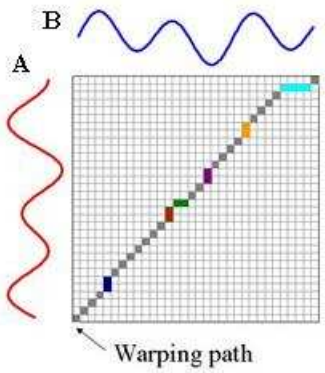

Fig. 2.2: Individuazione del *warping path* nell'algoritmo *Dynamic Time Warping*

I dati utilizzati spesso in letteratura, per esempio da Chu, Keogh, Hart, and Pazzani [2002], per mostrare la potenzialità di questo algoritmo sono i *Cylinder-Bell-Funnel* in Figura 2.3.

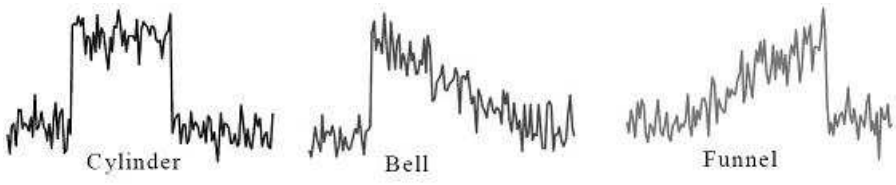

Fig. 2.3: Cylinder-Bell-Funnel

Questi dati simulati rappresentano tre tipologie di classi di serie e vengono generati dalle seguenti equazioni:

$$\begin{aligned}
c(t) &= (6+\eta)X_{[a,b]}(t) + \epsilon(t) \\
f(t) &= (6+\eta)X_{[a,b]}(t)\frac{(t-a)}{(b-a)} + \epsilon(t) \\
b(t) &= (6+\eta)X_{[a,b]}(t)\frac{(b-t)}{(b-a)} + \epsilon(t) \\
X_{[a,b]} &= \{1, if\ a \leq t \leq b,\ else\ 0\} \\
t &\in (1,100) \\
\eta &\sim N(0,1) \\
\epsilon &\sim N(0,1) \\
a &\sim U(16,32) \\
b &\sim U(32,96)
\end{aligned}$$

Le equazioni $c(t)$,$f(t)$ e $b(t)$ descrivono rispettivamente l'andamento delle serie appartenenti alla tipologia *Cylinder*,*Funnel* e *Bell*.
Si vuole quindi individuare un algoritmo di classificazione in grado di separare questo insieme di serie storiche nelle tre classi.
A titolo esemplificativo si può costruire un dataset costituito da 30 curve, 10 per ciascuna tipologia.
A ciascuna serie storica è stato associato un identificativo nel seguente modo:

- identificativo da 1 a 10 per la tipologia *Cylinder*
- identificativo da 11 a 20 per la tipologia *Bell*
- identificativo da 21 a 30 per la tipologia *Funnel*

In Figura 2.4 si possono vedere le serie simulate divise per tipologia. È stata quindi costruita la matrice delle distanze Euclidee e della distanza relativa all'algoritmo *Dynamic Time Warping*. Queste matrici sono state l'input dell'algoritmo gerarchico agglomerativo del legame medio che sulla base della

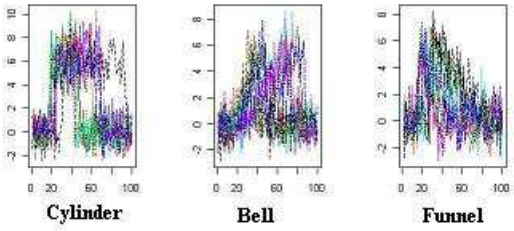

Fig. 2.4: Tre tipologie di serie storiche.

matrice di similarità (o di dissimilarità) aggrega le osservazioni più simili.
Dai dendrogramma in Figura 2.5 si può vedere come cambia la classificazione delle serie applicando il metodo di raggruppamento del legame medio alla matrice delle distanze Euclidee e alla matrice delle distanze *Dynamic Time Warping*. Nel secondo caso, tagliando il dendrogramma indicando un

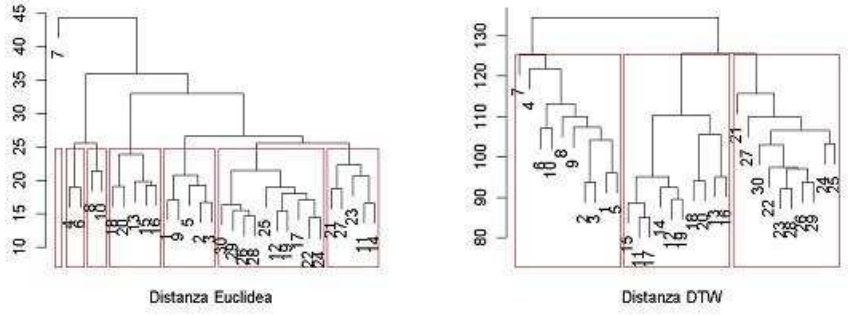

Fig. 2.5: Dendrogramma relativi al metodo del legame medio applicato a due tipologie di distanza

numero predefinito di gruppi pari a 3, si ha un raggruppamento perfetto delle serie storiche.
In questo caso si può vedere come con la distanza basata sul *Dynamic Time Warping* si riescono a classificare le serie in modo migliore rispetto alla distanza Euclidea.
Come di vedrà nel paragrafo successivo, applicando questa tecnica ai dati sul traffico telefonico non si riesce agevolmente ad individuare dei pattern tipici. Infatti queste serie sono caratterizzate da una forte variabilità e non ci sono delle conoscenze a priori che possono essere utilizzate per avere una prima idea dei possibili andamenti tipici.
Se la profondità temporale delle serie storiche a disposizione fosse stata più lunga si sarebbe potuto procedere ad un'analisi statistica delle serie storiche individuando per esempio la componente stagionale e utilizzare solo questa informazione come input dell'algoritmo di classificazione, riducendo la va-

riabilità delle osservazioni.
Spesso queste tecniche per confrontare e classificare le serie storiche vengono effettuate su serie che appartengono al mondo biomedico, per esempio con lo studio degli elettrocardiogramma, dei livelli ormonali o della manifestazione genica.
In tutti questi ambiti l'andamento delle serie storiche è stato studiato a sufficienza ed è ben noto che si possono ritrovare degli andamenti tipici.

2.3.1 Applicazione del *DTW* al traffico telefonico

Come accennato nel paragrafo precedente, con i dati disponibili, relativi al numero di telefonate effettuate e del numero di *sms* e di *mms* inviati, la ricerca di *pattern* utilizzando l'algoritmo *Dynamic Time Warping* non porta a dei buoni risultati.
In Figura 2.6 si riporta, a titolo esemplificativo, un sottoinsieme di serie dei dati oggetto di studio e su di esso si applica l'algoritmo *DTW*.
Queste serie storiche riassumono l'utilizzo del telefono cellulare da parte dei clienti di una compagnia telefonica, in cui ciascun punto rappresenta il numero complessivo di telefonate, di *sms* e di *mms* in uscita dalla singola *sim card*.
Applicando questo algoritmo ai dati a disposizione, ci si aspettava di individuare degli andamenti ben distinti nell'utilizzo del telefono cellulare, e in particolare, quegli andamenti di particolare interesse per l'azienda, quali per esempio i *trend* decrescenti.
Per facilitare l'algoritmo di ricerca dei pattern ho filtrato le serie eliminando quelle che presentano valori estremi, in particolare non sono state considerate le serie con punti con valori maggiori di 200. Inoltre è stata ridotta la dimensione temporale considerando solo i dati relativi ai mesi da Novembre 2004 ad Ottobre 2005, considerando la media aritmetica dei valori per i mesi per cui si avevano a disposizione più dati. Si sono considerate solo 50 serie storiche.
Da questo sottoinsieme di curve si riescono ad individuare diverse tipologie di andamenti. Per esempio alcune serie storiche sono caratterizzate da degli *spike*, mentre altre da dai livelli iniziali elevati per poi diminuire.
A partire da questo insieme di serie storiche spno state costruite le matrici basate sulla distanza Euclidee e sulla distanza *DTW*. Dopodichè si è proceduto a raggruppare queste serie secondo l'algoritmo del legame medio ottenendo i dendrogramma in Figura 2.7. Tagliando il dendrogramma in Figura 2.7 specificando un numero di gruppi pari a 5 si ottengono dei gruppi, riportati in Figura 2.8 e in Figura 2.9, che non presentano andamenti tipici e distintivi.
Si osserva che la distanza Euclidea non riesce a cogliere molte delle somiglianza che si riescono invece ad individuare con la distanza DTW. Infatti

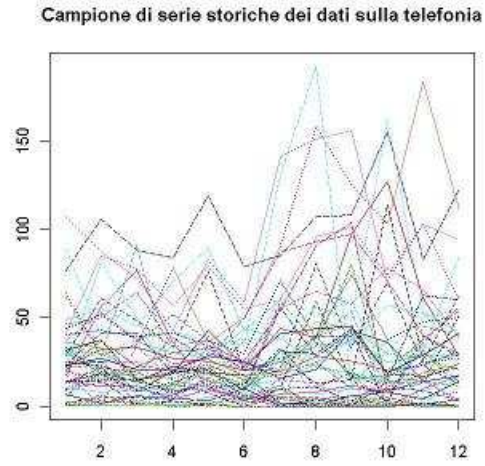

Fig. 2.6: Campione di serie storiche dei dati sulla telefonia

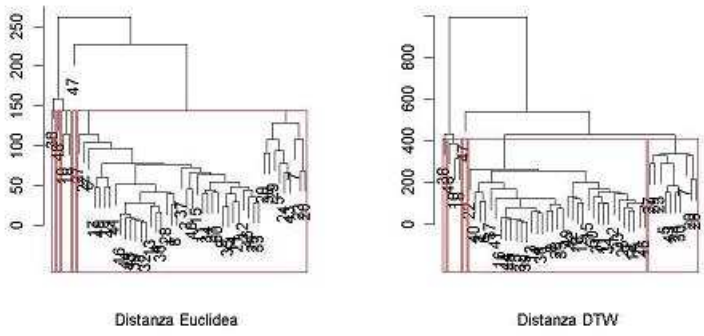

Fig. 2.7: Dendrogramma relativi al metodo del legame medio applicato a due tipologie di distanza

la distanza Euclidea individua un unico *elephant cluster*, in Gruppo con etichetta 5, e altri 4 gruppi composti da una o due serie storiche. Con la distanza DTW invece si individuano almeno 3 gruppi composti da un numero sufficientemente elevato di unità statistiche, e si nota che nel gruppo più popoloso non rientrano quelle serie storiche con valori del numero di telefonate effettuate e di messaggi inviati superiori ad 80 unità. Ma anche in questo caso i risultati sono poco soddisfacenti, in quanto non si riescono a caratterizzare agevolmente i gruppi individuati, in particolare per quanto riguarda il gruppo più numeroso che presenta serie storiche con *spike* sia nei primi che negli ultimi mesi di osservazione.

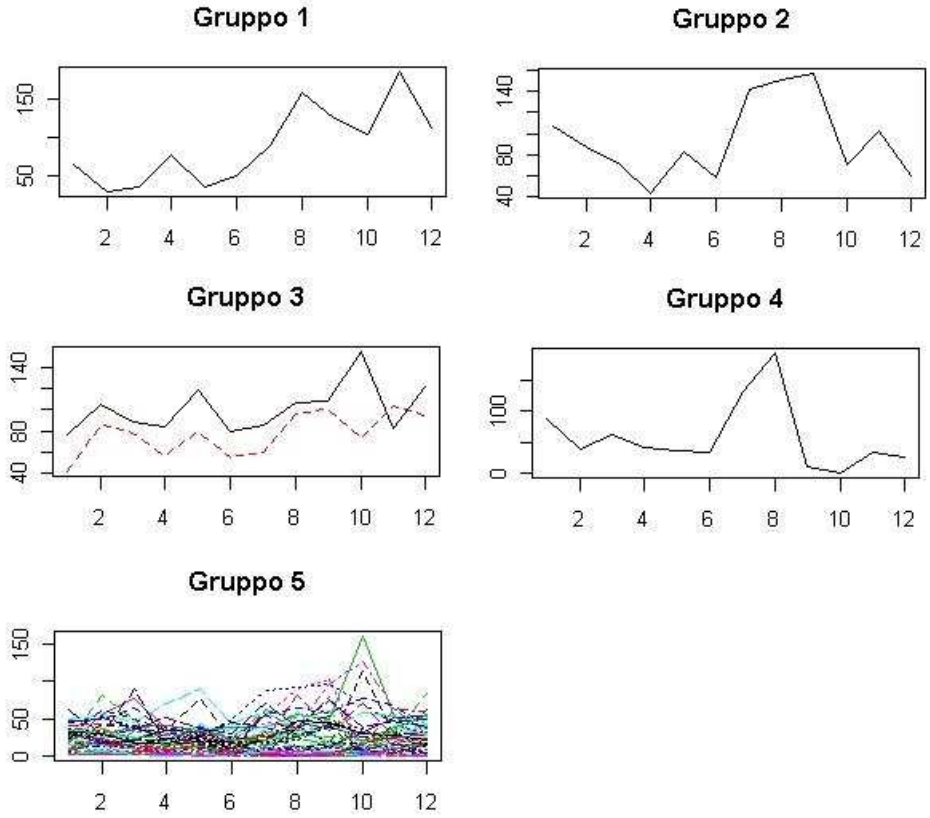

Fig. 2.8: Gruppi di serie storiche individuati con la distanza Euclidea

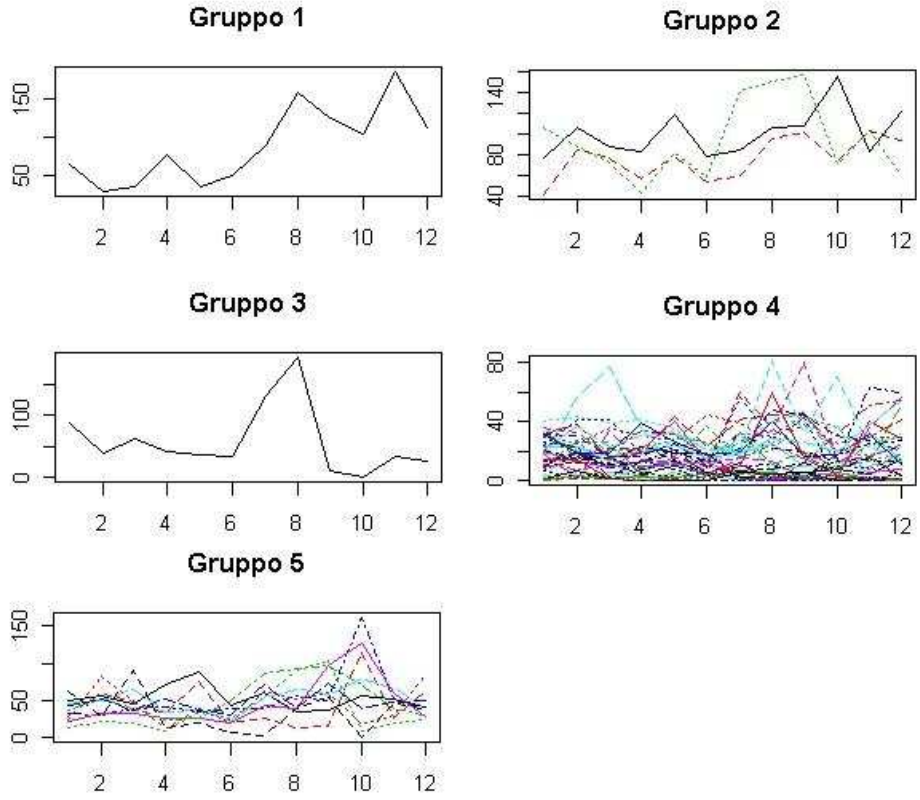

Fig. 2.9: Gruppi di serie storiche individuati con la distanza *Dynamic Time Warping*

Capitolo 3

Analisi funzionale

3.1 Dati Funzionali

Le serie storiche che rappresentano l'andamento del traffico telefonico verranno trattate e analizzate considerandole dei dati funzionali.
Quando si parla di dati funzionali ci si riferisce ad insiemi di dati collezionati per ciascuna unità statistica la cui caratteristica principale è che possono essere rappresentati come funzioni del tempo o dello spazio.
I dati funzionali si trovano in diversi campi, da quello biologico, come negli studi effettuati da Scarpa and Dunson [2009] e da Bigelow and B.Dunson [2007] che si occupano dell'andamento dei livelli ormonali, a quello economico con lo studio dell'andamento degli indici azionari, dell'andamento dei traffico telefonico,etc.
Tra i dati che rientrano nella definizione di dati funzionali troviamo:

- i dati longitudinali che possono essere interpolati per vedere come evolvono nel tempo.
  Per esempio si può considerare una popolazione e osservare per ciascun soggetto un gran numero di osservazioni e il cui dato funzionale associato è la stima della funzione di densità di queste osservazioni.
  Oppure, rivolgendoci al campo economico, si può osservare come varia nel tempo l'indice di determinati beni per individuarne eventuali dinamiche tipiche (Focardi [2001]).

- le curve tracciate nello spazio, che possono essere analizzate per studiare, per esempio, la calligrafia delle persone, registrando diverse volte delle parole scritte manualmente e vedere come variano nel tempo (Ramsay and Silverman [2002]).

- la forma di oggetti in due dimensioni, con i quali si può, per esempio, investigare la presenza di malformazioni ossee modellando la variabilità delle forme di queste all'interno di una popolazione, in modo da

cogliere le differenze tra ossa sane e malate (Ramsay and Silverman [2002]).

Gli obiettivi principali dell'analisi dei dati funzionali sono gli stessi delle altre metodologie statistiche i cui dati di input non sono funzionali, ma sono delle variabili indipendenti tra loro. Questi obiettivo sono:

- sviluppare strategie di presentazione dei dati in modo da sottolineare caratteristiche interessanti ed importanti;
- studiare la variabilità di un fenomeno e la presenza di trend ricorrenti;
- costruire modelli per i dati osservati per spiegare la dipendenza tra le variabili e tra le osservazioni.

Focalizzandoci sul secondo punto, quando in una stessa popolazione si osservano diversi andamenti di uno stesso fenomeno, quindi, quando ciò che differenzia due osservazioni è la forma stessa della distribuzione del dato funzionale associato all'osservazione, allora è opportuno procedere alla classificazione delle osservazioni in diverse tipologie di funzioni che possono essere più o meno simili.
Rivolgendoci ai metodi classici di classificazione gerarchici e non gerarchici, ci si accorge che non tengono in considerazione la correlazione dei punti nel tempo, caratteristica peculiare dei dati funzionali.
Si potrebbe ricorrere quindi a tecniche quali modelli normali multivariati per considerare la correlazione dei punti, ma questo tipo di modelli non tiene in considerazione l'ordine temporale.
In questi casi invece il tempo ha un ruolo importante nel discriminare tra una serie e l'altra.
È anche possibile utilizzare modelli autoregressivi per descrivere l'andamento di una serie storica, ma questi richiedono la stazionarietà della serie. Requisito non sempre soddisfatto dai fenomeni presi in considerazione.
Per questi motivi si è deciso di procedere all'analisi dei livelli del traffico telefonico secondo un'analisi funzionale dei dati che si presentano sotto forma di curve, cioè di dati funzionali.
Secondo questo approccio è possibile individuare delle possibili famiglie parametriche di dati longitudinali sfruttando, per esempio, una rappresentazione dei dati come combinazione lineare di basi di funzioni. In questo modo si definisce un modello parametrico sottostante i dati che, dall'analisi a posteriori dei parametri del modello, permette di poter individuare degli andamenti tipici nel fenomeno analizzato.
Per sfruttare eventuali informazioni a priori sul fenomeno, è inoltre possibile, secondo l'approccio bayesiano, assegnare delle distribuzione a priori ai parametri che definiscono il modello parametrico.
Si possono così implementare metodi di clustering Bayesiani che riescono a

combinare le informazioni a disposizione e a produrre una misurazione dell'incertezza per tutte le quantità stimate.

3.2 Analisi funzionale bayesiana

Per l'analisi dei dati funzionali sono stati proposti diversi approcci flessibili che permettono di procedere all'analisi dei dati pur non conoscendo né l'andamento medio né la distribuzione intorno alla media.
Con lo scopo di modellare l'andamento del traffico telefonico, questo lavoro, sulla base della metodologia proposta da Scarpa and Dunson [2009], applica un approccio flessibile che incorpora le informazioni a priori in un'analisi funzionale Bayesiana semiparmetrica.
L'approccio proposto è basato sulla specificazione della distribuzione delle funzioni come miscuglio di un modello parametrico e di una componente non parametrica. La componente parametrica è stata formulata sulla base di motivazioni di marketing mentre la componente non parametrica è caratterizzata da un processo funzionale di Dirichlet.
La componente non parametrica consente di individuare andamenti più o meno marginali nell'andamento del traffico telefonico. I modelli sono stati sviluppati per una stima della distribuzione a posteriori e sono stati condotti su dati relativi alla *Customer Base* di una compagnia telefonica.

3.3 Costruzione del modello miscuglio bayesiano

Come introdotto nel Capitolo 1, lo studio della distribuzione di una curva casuale è un tema d'interesse in molte applicazioni.
Sulla base delle conoscenze a priori che si hanno sul fenomeno si ipotizzano, o in alcuni casi sono noti, gli andamenti tipici dei dati longitudinali in esame. Questi andamenti caratteristici vengono quindi usati come base dell'algoritmo per identificare le curve che seguono un andamento normale da quelle affette da comportamenti anomali.
Per caratterizzare le curve dei diversi soggetti si può usare un modello della forma:

$$\begin{aligned} y_i(t) &= \eta_i(t) + \varepsilon_i(t) \\ \varepsilon_i(t) &\sim N(0, \sigma^2) \\ \sigma^{-2} &\sim Gamma(a, b) \\ \eta_i &\sim G_i \\ \mathcal{G} = \{G_i\}_{i=1}^n &\sim P \end{aligned} \tag{3.1}$$

dove:

$t = 1, \ldots, T$	istante temporale
$i = 1, ..., n$	soggetti
$y_i(t)$	livello del traffico telefonico all'istante t dell'*i-esimo* soggetto
$\eta_i : T \longrightarrow \Re$	funzione smooth con $T \subseteq \Re^+$
$\varepsilon_i(t)$	componente erratica
(a, b)	parametri prespecificati
G_i	distribuzione a priori delle distribuzioni per l'$i - esimo$ soggetto
$\mathcal{G}$	insieme delle distribuzioni delle funzioni dei diversi soggetti
P	distribuzione a priori per $\mathcal{G}$

L'espressione (3.1) fornisce una struttura per l'analisi dei dati funzionali le cui componenti possono essere specificate tramite diversi approcci.
Una stategia comune è quella di semplificare il problema considerando una rappresentazione lineare in basi:

$$\eta_i(t) = \sum_{h=1}^{q} \theta_{i,h} b_h(t), \forall t \in \mathcal{T} \tag{3.2}$$

utilizzando quindi una base di q funzioni, dove:

$\{b_h\}_{h=1}^q$	base di funzioni pre-assegnata
$\theta_i = [\theta_{i,1}, \ldots, \theta_{i,q}]'$	coefficienti delle basi relative alle osservazioni

Si deve quindi individuare la distribuzione a priori di questi q coefficienti, per i quali si può scegliere un modello normale :

$$\begin{aligned}
\theta_i &\sim N_q(\alpha_i, \Omega) \\
\alpha_i &\sim N_q(\alpha, \Omega_1) \\
\alpha &\sim N_q(\alpha_0, \Omega_0) \\
\Omega &= diag\{\omega_1, \omega_2\} \\
\Omega_1 &= diag\{\omega_{1,1}, \omega_{1,2}\} \\
\omega_1 &\sim Gamma(c, d) \\
\omega_2 &\sim Gamma(e, f) \\
\omega_{1,1} &\sim Gamma(g, h) \\
\omega_{1,2} &\sim Gamma(i, l)
\end{aligned} \tag{3.3}$$

dove:

$\alpha_i =$	media dei coefficienti delle basi soggetto specifici
$\alpha =$	media dei coefficienti delle basi complessiva
$\Omega_1 =$	misura la variabilità della media per ciascun soggetto
$\Omega =$	misura la variabilità della media complessiva

e dove $(\alpha_0, \Omega_0, c, d, e, f, g, h, i, l)$ sono iper parametri prespecificati.

In letteratura vengono proposte varianti a questa formulazione:

- Bigelow and B.Dunson [2007] e Thompson and Rosen [2007] propongono di aggiungere incertezza nella selezione delle funzioni delle basi.
- Ray and Mallick [2006], per evitare l'ipotesi di normalità dei coefficienti delle basi all'interno del modello, propongono di usare come a priori un Processo di Dirichlet(DP) (Ferguson [1973]).
- Rodriguez, Dunson, and Gelfand [2009]) evitano la rappresentazione delle basi di funzioni attraverso l'utilizzo di misture DP.

Dato il carattere quasi certamente discreto del DP, questi approcci raggruppano i soggetti in classi funzionali.
Inserendo nel modello anche una componente gerarchica, si possono considerare i lavori di:

- Scarpa and Dunson [2009] che si propongono di modellare i livelli del *basal body temperature*,*bbt* analizzando i dati longitudinali di una popolazione di donne in età feconda. La struttura gerarchica che gli autori hanno tenuto in considerazione nella costruzione del modello è definita dall'insieme delle donne e dall'insieme delle curve del *bbt* relativo a ciascuna donna.
- Teh, Jordan, Beal, and Blei [2006] che con l'obiettivo di generalizzare questi metodi con un'impostazione a due livelli, utilizzano come a priori per $\mathcal{G}$ un DP Gerarchico (HDP).
- Heard, Holmes, and Stephens [2006] che propongono un approccio Bayesiano per la classificazione di funzioni gerarchiche.

Questi approcci sono utili nel caso in cui ci siano poche informazioni a priori circa la forma delle funzioni in studio e si vuole un modello altamente flessibile. Infatti anche nel caso in cui un modello parametrico noto riuscirebbe a fornire una buona descrizione della maggior parte delle curve, accade spesso che un sottoinsieme di queste curve sia irregolare.
In questi casi si riesce ad individuare un modello parametrico per modellare gli andamenti tipici, ma rimane difficile parametrizzare tutte le possibili *pattern*.
Per risolvere questo problema, in questo elaborato si decide, quindi, di modellare $\mathcal{G}$ usando una miscuglio di una componente parametrica con una contaminazione non parametrica. In particolare la componente non parametrica sarà caratterizzata da un DP funzionale.

3.3.1 Componente parametrica

Con l'obiettivo di modellare le curva, si propone una funzione lineare a tratti con tre semplici componenti che riassumono l'andamento tipico del traffico telefonico:

1. una prima linea orizzontale alta che descrive il *plateau* iniziale
2. una linea intermedia obliqua descrive la velocità nel decremento
3. una seconda linea orizzontale bassa che descrive il *plateau* finale.

Con questa struttura lineare a tratta e permettendo ai parametri di aggiornarsi sulla base delle dati osservati è possibile modellare contemporaneamente gli andamenti caratterizzati da cali più o meno veloci con la presenza eventuale di *plateau*, di ampiezza e livello variabile, all'inizio e alla fine del periodo di osservazione.
Considerando la notazione introdotta nel Paragrafo 3.3, il modello statistico per il traffico telefonico è:

$$y_i(t) = \begin{cases} \theta_{i,1} + \varepsilon_i(t) & \text{se} \quad 1 \leq t \leq k_i \\ \theta_{i,1} + \theta_{i,2}(\frac{t-k_i}{r_i}) + \varepsilon_i(t) & \text{se} \quad k_i \leq t \leq k_i + r_i \\ \theta_{i,1} + \theta_{i,2} + \varepsilon_i(t) & \text{se} \quad k_i + r_i \leq t \leq T \end{cases}$$

$\varepsilon_{ij}(t) \sim N(0, \sigma^2)$ si considera quindi una base con $q = 2$ funzioni dove:

k_i	ultimo mese della fase iniziale
r_i	numero di mesi in cui il livello del traffico telefonico decresce
$\theta_{i,1}$	livello iniziale del traffico telefonico
$\theta_{i,2}$	tasso di decremento del traffico telefonico
σ^2	varianza della componente erratica

Questo modello può essere riscritto in forma matriciale come un modello lineare:

$$\begin{aligned} y_i &= X_i\theta_i + \varepsilon_i \\ \varepsilon_i &\sim N_{n_i}(0, I_{n_i}\sigma^2) \end{aligned} \tag{3.4}$$

dove:

$\theta_i = [\theta_{i,1}, \theta_{i,2}]'$

$$X_i = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & \frac{1}{r_i} \\ 1 & \frac{2}{r_i} \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} k_i \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} r_i \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} T - k_i - r_i \end{matrix}$$

Sotto il Modello (3.4) l'andamento del traffico è caratterizzato da:

$$\eta_i(t) = X_i(t)'\theta_i$$

dove:

$X_i(t)$ vettore delle basi al tempo t
θ_i vettore dei coefficienti delle basi

È immediato modificare il modello considerando altre basi con una formulazione più complessa.
Una specificazione bayesiana completa comprende le seguenti distribuzioni a priori:

$$\begin{aligned} k_i &\sim U(1,14) \\ r_i &\sim U(1,15) \end{aligned}$$

dove:

$U(a,b)$ indica una distribuzione Uniforme sull'intervallo (a,b)

3.3.2 Componente non parametrica e Processo di Dirichlet funzionale

In concomitanza al modello parametrico proposto nel paragrafo precedente, si propone un approccio bayesiano non parametrico per dei dati funzionali, per poter individuare i diversi pattern presenti nella popolazione. Si userà un Processo di Dirichlet Funzionale per permettere ad un sottoinsieme di curve di scostarsi dal modello parametrico, considerando quindi una contaminazione non parametrica.
Poiché l'obiettivo è quello di usare una contaminazione non parametrica per caratterizzare i diversi sottoinsiemi di traiettorie anomale, non si considererà la dipendenza nelle osservazioni che verranno classificate nella componente non parametrica.
Come approccio generale per caratterizzare le curve anomale, si raccomanda di scegliere un Processo Gaussiano, $GP(\mu, \mathcal{C})$, così che le realizzazioni corrispondano ad una grande varietà di plausibili forme a priori. In particolare è la funzione di covarianza $\mathcal{C}$ a controllare i tipi di curve osservati. Come semplice scelta standard, si consiglia una funzione di covarianza esponenziale, in modo da permettere una grande varietà di forme funzionali. Approcci più flessibili prevedono l'utilizzo di forme funzionali parametriche i cui parametri devono essere stimati sulla base dei dati (nel nostro caso pochi) del sottoinsieme di funzioni appartenenti alla componente non parametrica.
Supponiamo ora che:

$$\eta_i \sim G$$

dove:

η_i	è una rappresentazione *smooth* di ciascuna serie storica i
G	distribuzione non nota delle curve *smooth* tra i soggetti

Si assuma che:

$$\begin{aligned} \eta_i &\sim G = \sum_{h=1}^{\infty} p_h \delta_{\Theta_h} \\ \Theta_h &\sim GP(\mu, \mathcal{C}) \end{aligned}$$

dove:

$h = 1, ..., \infty$	numero di possibili classi non parametriche
Θ_h	insieme di elementi casuali i.i.d secondo un Processo Stocastico Gaussiano con funzione della media μ e funzione di covarianza $\mathcal{C}$
δ_{Θ_h}	è un elemento di Θ_h
$p_h = V_h \prod_{l=1}^{h}(1 - V_l)$	$h = 1, \ldots, \infty$
$V_h \sim Beta(1, \alpha_{Dir})$	$h = 1, \ldots, \infty$
$\alpha_{Dir} \sim Gamma(a, b)$	Parametro di precisione del Processo di Dirichlet

Si usa una rappresentazione *stick breaking* della misura di probabilità di G in accordo con la rappresentazione di un Processo di Dirichlet. L'espressione (3.5) descrive un Processo di Dirichlet Funzionale (FDP) per la distribuzione ignota delle curve G.
Sebbene G sia formalmente una misura di probabilità con supporto in uno spazio funzionale, si può concettualmente vedere G come caratterizzazione della distribuzione delle curve tramite una specificazione discreta che associa una massa di probabilità alle diverse forme, conducendo ad una classificazione funzionale.
In particolare se:

$$\eta_i \sim G$$

la probabilità a priori di allocazione dell'osservazione i nel cluster funzionale h è:

$$Pr[\eta_i = \Theta_h] = p_h$$

Il Processo Gaussiano induce una distribuzione normale multipla sui valori osservati di Θ_h.
Il parametro di precisione α_{Dir} del Processo di Dirichlet è stimato assumendo una distribuzione iper a priori Gamma.

3.3.3 Modello miscuglio

Al fine di incorporare le conoscenze a priori nell'andamento dei dati e la flessibilità del Processo Gaussiano, si propone di combinare il modello parametrico con il Processo di Dirichlet Funzionale attraverso un modello miscuglio.
Il modello miscuglio proposto è quindi:

$$\eta_i \sim \pi\delta_{X_i}\Theta_i + (1-\pi)G \tag{3.5}$$

dove:

$\delta_{X_i}\Theta_i$	rappresenta la componente parametrica
G	rappresenta la componente non parametrica

π rappresenta quindi la probabilità che la serie storica venga collocata nella componente parametrica.
Per semplicità si considera π fisso per tutte le serie storiche, ma consideriamo l'incertezza di questo parametro considerando:

$$\pi \sim Beta(\alpha, \beta)$$

dove α e β sono iper parametri pre specificati.
Si osserva che tutte le serie storiche classificate nella componente parametrica avranno un andamento del traffico telefonico che sarà ben descritto dal modello lineare a tratti.
Si hanno poche conoscenze a priori sui tipi di curve che descrivono le traiettorie che si discostano dal modello parametrico, ma grazie alla flessibilità della contaminazione non parametrica si riuscirà ad ottenere informazioni su questi andamenti.

3.3.4 Calcolo delle distribuzioni a posteriori

Il calcolo delle distribuzioni a posteriori procede attraverso un'immediata modifica dell'algoritmo *Polya Urn Gibbs Sampling* [1] largamente usato nei modelli miscuglio con un'apriori caratterizzata da un Processo di Dirichlet. In particolare l'algoritmo è strutturato nel seguente modo:

1. passo di *data augmentation* [2] per allocare i soggetti nella componente parametrica o non parametrica

[1] L'algoritmo Gibbs Sampling ha senso solo nel caso multidimensionale, in cui si ha una variabile casuale n-dimensionale, $X = (X_1, X_2, ..., X_n)$, le cui componenti X_i vengono simulate condizionatamente a tutte le altre.

[2] Il *Data Augmentation* è un algoritmo per la classificazione bayesiana utilizzato per ottenere un campione da una distribuzione a priori

2. passo di *Gibbs Sampler* per aggiornare le componenti non note del modello parametrico a partire dalla distribuzione condizionata

3. passo del *Polya Urn Gibbs Sampling* per aggiornare le componenti non note del modello non parametrico

Per la serie storica i si pone:

$S_i = -1$	se la serie storica viene allocata nella componente parametrica
$S_i = h$	se la serie storica viene allocata nel gruppo h della componente non parametrica
$S_i = 0$	se la serie storica viene allocata in un nuovo gruppo della componente non parametrica
$h = (1, .., s)$	possibili gruppi non parametrici

L'apice (i) indica un vettore ottenuto escludendo le osservazioni della serie storica i.
Gli *step* sono descritti di seguito.
Step 0 : Inizializzazione
L'algoritmo parte con tutte le curve nella componente parametrica, quindi con $s = 0$.
I parametri k_i e r_i vengono inizializzati ad un unico valore per tutte le serie storiche, mentre gli altri parametri della componente parametrica, i parametri della componente non parametrica e i parametri comuni vengono inizializzati estraendoli dalla loro distribuzione a priori.

Step 1 : Aggiornamento indicatori dei cluster

Per ogni serie storica, si estrae l'indicatore del cluster $S_i^{(i)}$ dalla distribuzione a posteriori multinomiale con probabilità:

$$Pr[S_i = h| \ldots] = q_{i,h}, h = -1, 0, 1, \ldots, s^{(i)}$$

dove:

$S_i^{(i)} = 0$	indica che la serie storica proviene da un nuovo cluster non parametrico

Per la stima del parametro $q_{i,h}$ si devono distinguere i tre casi in cui la serie storica viene collocata nella componente parametrica in cui $h = -1$, in nuova componente non parametrica $h = 0$ oppure in una componente parametrica già esistente $h > 0$.
1. Per h=-1

Considerando che il modello parametrico può essere scritto come un modello lineare gerarchico, si ottiene:

$$q_{i,-1} \propto c\pi \int N_{n_i}(y_i; X_i\theta_i, \sigma^2 I_T) N_T(\theta_i; \alpha_i, \Omega)\, d\theta_i$$

dove $N_n(.; \mu, \Sigma)$ è una densità normale multidimensionale con media μ e matrice di varianze e covarianze Σ.
Risolvendo l'integrale si ottiene la seguente distribuzione normale:

$$N(X_i\alpha_i, \sigma^2 I_T + X_i\Omega X_i^T)$$

Infatti si può notare che, date le seguenti distribuzioni:

$$\begin{aligned} y_i &\sim N_T(X_i\theta_i, \sigma^2 I_T) \\ \theta_i &\sim N_T(\alpha_i, \Omega) \end{aligned}$$

da cui deriva che:

$$X_i\theta_i \sim N_T(X_i\alpha_i, X_i\Omega X_i^T)$$

allora marginalmente y_i ha la seguente distribuzione:

$$y_i \sim N_T(X_i\alpha_i, \sigma^2 I_T + X_i\Omega X_i^T)$$

2. Per h=0

Considerando che l'andamento medio del traffico telefonico relativo alla componente non parametrica viene generato a partire da un Processo Gaussiano di media μ_0 e struttura di covarianza $\mathcal{C}$ si ottiene:

$$q_{i,0} \propto c(1-\pi)\alpha_{dir} \int N_T(y_i; \mu, \sigma^2 I_T) N_T(\mu; \mu_0, \mathcal{C})\, d\mu$$

Risolvendo l'integrale si ottiene la seguente distribuzione Normale multivariata:

$$N_T(\mu_0, \sigma^2 I_T + \mathcal{C})$$

Infatti si può notare che:

$$\begin{aligned} y_i &\sim N_T(\mu, \sigma^2 I_T) \\ \mu &\sim N_T(\mu_0, \mathcal{C}) \end{aligned}$$

Allora marginalmente, la distribuzione di y_i è:

$$y_i \sim N_T(\mu_0, \sigma^2 I_T + \mathcal{C})$$

3. Per $h > 0$

L'algoritmo, sfruttando le informazioni ottenute sugli h gruppi che si sono formati, alloca le singole osservazioni nei cluster indicati da h:

$$q_{i,h} \propto c(1-\pi)n_h^{(i)} N_T(y_i; \psi_h^{(i)}, \sigma^2 I_T)$$

dove:

$n_h^{(i)}$ è il numero di curve nel gruppo h senza considerare l'$i-esima$ curva;
$\psi_h^{(i)}$ è il vettore delle medie del gruppo h senza considerare l'$i-esima$ curva;

Per ogni i tale che $S_i^{(i)} = 0$, s viene aggiornato con $s = s + 1$ e viene estratto un nuovo vettore delle medie del cluster, ψ_h, la cui distribuzione a posteriori è:

$$[\psi_s | \ldots] \sim N(\psi_s; \mu_0, C_T) N_T(y_i; \psi_s, \sigma^2 I_T)$$

dove:

$N(\psi_s; \mu_{0,i}, C_T)$ è la distribuzione a priori di ψ_s
$N_T(y_i; \psi_s, \sigma^2 I_T)$ è la verosimiglianza dell'osservazione y_i

Step 2: Aggiornamento del modello parametrico

In questo *step*, dopo avere allocato le curve nella componente parametrica o nelle h componenti non parametriche, vengono aggiornati i parametri relativi alla componente parametrica.
Per quanto riguarda l'aggiornamento di k_i e r_i che indicano la fase di incremento o decremento dei livelli del traffico telefonico, si procede al confronto della densità associata a ciascuna osservazione in corrispondenza dei possibili valori di k_i e r_i. In pratica si fissa il valore di k_i individuato nell'iterazione precedente e si va a trovare quindi una stima di massima verosimiglianza per r_i, e si fa lo stesso per r_i fissando il valore di k_i relativo all'iterazione precedente.

$$\widehat{k}_i = \max_{k_i} N_T(y_i, X_i^k \theta_i, \sigma^2 I_T)$$
$$\widehat{r}_i = \max_{r_i} N_T(y_i, X_i^r \theta_i, \sigma^2 I_T)$$

dove, la matrice delle basi X_i^r è definita dalla coppia (k_i, r_i) con r_i fissato e la matrice X_i^k è definita dalla coppia (k_i, r_i) con k_i fissato e i parametri θ_i e σ^2 sono relativi all'iterazione precedente.

Per quanto riguarda invece l'aggiornamento dei parametri θ_i e σ_2 del modello parametrico, si applica un algoritmo di simulazione *Markov Chain Monte Carlo*, il *Gibbs Sampling*, al modello gerarchico lineare, in cui solo le curve allocate nel modello parametrico verranno utilizzate.

In questo caso abbiamo delle osservazioni $X \sim N(\mu, \sigma^2)$ dove μ e σ^2 dipendono da parametri di cui abbiamo una distribuzione a priori.
Considerando quindi il seguente modello gerarchico:

$$\begin{aligned} y_i &= X_i\theta_i + \varepsilon_i \\ \varepsilon_i &\sim N_T(0, I_T\sigma^2) \end{aligned}$$

con le seguenti distribuzioni a priori:

$$\begin{aligned} \theta_i &\sim N_q(\alpha_i, \Omega_1) \\ \alpha_i &\sim N_q(\alpha, \Omega) \\ \alpha &\sim N_q(\alpha_0, \Omega_0) \end{aligned}$$

Indicando con $n_T = \sum_{i=1}^{n} n_i$ il numero totale delle curve assegnate alla componente parametrica, l'algoritmo di *Gibbs Sampling* per la stima dei parametri estrae iterativamente dalle seguenti distribuzioni *full conditional*:

$$\begin{aligned} \theta_i|y_i &\sim N_q(V_1 b_1, V_1) \\ V_1 &= (\Omega_1^{-1} + \sigma^{-2} X_i^T X_i)^{-1} \\ b_1 &= (\Omega_1^{-1}\alpha_i + \sigma^{-2} X_i^T y_i) \end{aligned}$$

da questo *step* si ottiene la matrice θ_i con q colonne che rappresentano i coefficienti delle basi e n_T righe.

$$\begin{aligned} \alpha_i|\theta_i, y_i &\sim N_q(V_2 b_2, V_2) \\ V_2 &= (\Omega^{-1} + \Omega_1^{-1} n_T)^{-1} \\ b_2 &= (\Omega^{-1}\alpha + \Omega_1^{-1} \sum_{i=1}^{n_T} \theta_i) \end{aligned}$$

da questo *step* si ottiene la matrice α_i con q colonne per n_T righe.

$$\begin{aligned} \alpha|\theta_i, y_i &\sim N_q(V_3 b_3, V_3) \\ V_3 &= (\Omega_0^{-1} + \Omega^{-1} n_T)^{-1} \\ b_3 &= (\Omega_0^{-1}\alpha_0 + \Omega^{-1} \sum_{i=1}^{n_T} \alpha_i) \end{aligned}$$

da questo *step* si ottiene il vettore α di q elementi.

$$\begin{aligned} \omega_{1,l}^{-1} &\sim \Gamma(a_{1,2}, b_{1,2}) \\ a_{1,2} &= a_{1,l} + n_T/2 \\ b_{1,2} &= b_{1,l} + 0.5 \sum_{i=1}^{n_T} (\theta_{i,l} - \alpha_{i,l})^2 \\ l &= 1, \ldots, q \end{aligned}$$

da questo *step* si ottiene il vettore ω_1 di q componenti.

Infatti dato che:

$$\begin{aligned} \theta_i &\sim N_q(\alpha, \Sigma) \\ \Sigma &= \begin{bmatrix} \omega_1 & \ldots & 0 \\ \vdots & \omega_l & \vdots \\ 0 & \ldots & \omega_q \end{bmatrix} \end{aligned}$$

in cui ciascun elemento della matrice Σ ha una distribuzione a priori *Gamma*:

$$\omega_{1,l} \sim Gamma(a_{1,l}, b_{1,l}), l = 1, \ldots, q$$

Per calcolare la distribuzione a posteriori di ciascuno iper parametro della distribuzione *Gamma* devo considerare la verosimiglianza marginale di ciascuna componente del vettore di parametri θ_i:

$$\begin{aligned} \theta_i &= [\theta_{i,1}, \ldots, \theta_{i,q}]^T \\ \alpha_i &= [\alpha_{i,1}, \ldots, \alpha_{i,q}]^T \\ \theta_{i,l} \sim N(\alpha_{i,l}, \omega_{1,l}) & \\ l &= 1, \ldots, q \end{aligned}$$

Quindi considerando che per ciascun parametro $\theta_{i,l}$ ho un campione di ampiezza n_T, la distribuzione a posteriori di ω_l sarà una *Gamma* i cui nuovi parametri riceveranno un contributo da ciascuna di queste osservazioni.
Analogamente a ω_1 si ottiene il vettore ω:

$$\begin{aligned} \omega_l^{-1} &\sim \Gamma(a_2, b_2) \\ a_2 &= a_l + n_T/2 \\ b_2 &= b_l + 0.5 \sum_{i=1}^{n_T} (\alpha_{i,l} - \alpha_l)^2 \\ l &= 1, \ldots, q \end{aligned}$$

da questo *step* si ottiene il vettore ω di q componenti.

Step 3 : Aggiornamento componenti non parametriche

Aggiorna i parametri specifici dei cluster usando solo i soggetti di quel cluster.
Si indica con $\psi_i(t)$ il valore di ψ_i al tempo t.
La verosimiglianza per i soggetti nei cluster indicati da h della componente non parametrica è:

$$\prod_{i:S_i=h} N_T(y_i; \psi_h(t), \sigma^2 I_T)$$

La distribuzione a posteriori per il valore di ψ_h valutato in ogni collezione finita di punti t è proporzionale a questa verosimiglianza moltiplicata per un'a priori normale per $\psi_h(t)$ che deriva del Processo Gaussiano a priori $GP(\psi_h; \mu, C)$.
Dato che, nel caso del modello non parametrico, ciascuna variabile casuale y_i è distribuita secondo una Normale di media ψ_h e varianza $\sigma^2 I_T$:

$$y_i \sim N_T(\psi_h(t), \sigma^2 I_T)$$

per semplicità di notazione pongo:

$$\psi_h(t) = \mu$$
$$\sigma^2 I_T = SS$$

e dato che a priori ψ_h si distribuisce secondo una Normale di media μ e varianza C:

$$\psi_h \sim N_T(\mu, C)$$

per semplicità di notazione pongo:

$$\mu = \mu_0$$
$$C = S_0$$

si ottiene come distribuzione a posteriori:

$$\psi_h \sim N_T(\mu^*, \Sigma^*)$$

dove:

$$\mu^* = S_0(S_0 + \frac{1}{n_h}SS)^{-1}\bar{y}_i + \frac{1}{n_h}SS(S_0 + \frac{1}{n_h}S)^{-1}\mu_0$$
$$\sigma^* = S_0(S_0 + \frac{1}{n_h}SS)^{-1}\frac{1}{n_h}SS$$

Per quanto riguarda l'aggiornamento di α_{dir}, il parametro di precisione del Processo di Dirichlet, sulla base del lavoro svolto da Escobar and West [1995], si decide di incorporarlo all'interno dell'algoritmo *Gibbs Sampling*. A questo punto dell'analisi abbiamo a disposizione s il numero di possibili componenti non parametriche.
La distribuzione a priori di α_{dir} sappiamo essere una $Gamma(a, b)$ i cui parametri a e b sono degli iper-parametri pre specificati.
Per poter calcolare la distribuzione a posteriori di α_{dir} dobbiamo specificare una distribuzione a priori anche per il numero di possibili componenti parametriche s che dipende da α_{dir} e dalla numerosità campionaria n_h:

$$P(s|\alpha_{dir}, n_h) = c_{n_n}(s) n_h! \alpha_{dir}^s \Gamma[\alpha_{dir}]/\Gamma[\alpha_{dir} + n_h]$$

eliminando le componenti che non dipendono da α_{dir} otteniamo:

$$P(s|\alpha_{dir}) = \alpha_{dir}^s Gamma[\alpha_{dir}]/\Gamma[\alpha_{dir} + n_h]$$

Quindi la distribuzione a posteriori per α_{dir} è data da:

$$p(\alpha_{dir}|s, y_i) = p(\alpha_{dir}|s)\ p(\alpha_{dir} p(s|\alpha_{dir})$$

L'algoritmo di *Gibbs Sampling* anche in questo caso procede nel seguente modo:

1. si parte da un valore iniziale per α_{dir}
2. si estraggono i parametri del modello v dalla loro distribuzione condizionata $p(v|\alpha_{dir}, y_i)$
3. si estrae s
4. si estrae un nuovo valore di α_{dir} dalla distribuzione a posteriori $p(\alpha_{dir}|s, v, y_i)$

In questi casi, quando si estraggono dei valori da una distribuzione a posteriori, si possono utilizzare dei metodi per accettare o meno il valore proposto (Metodo di Accettazione-Rifiuto) in relazione alla forma della distribuzione a priori di $p(\alpha_{dir})$.
Estrarre da una distribuzione a posteriori continua esatta, è possibile con l'algoritmo iterativo *Gibbs Sampling* quando l'a priori di α_{dir} proviene da una classe di miscuglio di distribuzioni Gamma, o più semplicemente da una distribuzione Gamma.
Supponiamo quindi che :

$$\alpha_{dir} \sim Gamma(a, b)$$

In questo caso abbiamo che la distribuzione a posteriori è una miscuglio di due Gamma, e la distribuzione del parametro di miscuglio, dato α_{dir}, s e

n_h, ha una distribuzione $Beta$.
Per α_{dir} >0 la funzione $\Gamma(.)$ della distribuzione a priori di s può essere riscritta come:

$$\Gamma(\alpha_{dir})/\Gamma(\alpha_{dir}+n_h) = (\alpha_{dir}+n_h)Beta(\alpha_{dir}+1, n_h)/(\alpha_{dir}\Gamma(n_h))$$

quindi possiamo riscrivere la distribuzione a posteriori di α_{dir}, per $s > 1$ intero:

$$p(\alpha_{dir}|s) \sim p(\alpha_{dir})\alpha_{dir}^{s-1}(\alpha_{dir}+n_h)Beta(\alpha_{dir}+1, n_h)$$

secondo la definizione della funzione $Beta$ possiamo riscrivere:

$$p(\alpha_{dir}|s) \sim p(\alpha_{dir})\alpha_{dir}^{s-1}(\alpha_{dir}+n_h)\int_0^1 (x\alpha_{dir}(1-x)^{n_h-1})dx$$

Possiamo quindi affermare che $p(\alpha_{dir}|s)$ è la distribuzione marginale di una unione tra α_{dir} e una quantità continua η tale che:

$$p(\alpha_{dir}, \eta|s) \sim p(\alpha_{dir})\alpha_{dir}^{s-1}(\alpha_{dir}+n_h)\eta^{\alpha_{dir}}(1-\eta)^{n_h-1}$$

dove:

$$\begin{aligned} \alpha_{dir} &> 0 \\ 0 &< \eta < 1 \end{aligned}$$

Quindi abbiamo le seguenti distribuzioni a posteriori:

$$p(\alpha_{dir}|\eta, s) \sim \alpha_{dir}^{a+s-2}(\alpha_{dir}+n_h)exp(-\alpha_{dir}(b-log(\eta)))$$

possiamo riscrivere:

$$p(\alpha_{dir}|\eta, s) \sim \alpha_{dir}^{a+s-1}exp(-\alpha_{dir}(b-log(\eta))) + n_h\alpha_{dir}^{a+s-2}exp(-\alpha_{dir}(b-log(\eta)))$$

con:

$$\alpha_{dir} > 0$$

Quindi abbiamo che la distribuzione a posteriori di α_{dir} è definita tramite la seguente distribuzione miscuglio:

$$(\alpha_{dir}|\eta, s) \sim \pi_\eta\Gamma(a+s, b-log(\eta)) + (1-\pi_\eta) * \Gamma(a+s-1, b-log(\eta))$$

dove i pesi π_η sono così definiti:

$$\pi_\eta/(1-\pi_\eta) = (a+s-1)/(n_h(b-log(\eta)))$$

Queste distribuzioni sono ben definite per:

- ogni a priori di tipo *Gamma*
- ogni η definito in un intervallo unitario
- ogni $s > 1$

$$p(\eta|\alpha_{dir}, s) \sim \eta^{\alpha_{dir}}(1-\eta)^{n_h-1}$$

dove:

$$0 < \eta < 1$$

quindi η ha una distribuzione *Beta*:

$$(\eta|\alpha_{dir}, s) \sim Beta(\alpha_{dir}+1, n_h)$$

con media $(\alpha_{dir}+1)/(\alpha_{dir}+n_h+1)$.
Quindi ad ogni iterazione del *Gibbs Sampling* i valori correnti di s e α_{dir} permettono di estrarre un nuovo valore di α_{dir}, nel seguente modo:

1. si estrae un nuovo valore di η dalla distribuzione *Beta* (condizionatamente ai valori di α_{dir} e s più recenti)
2. si estrae α_{dir} dal modello miscuglio (utilizzando lo stesso s dello *step* precedente e utilizzando il nuovo valore di η)

Alla fine della simulazione, $p(\alpha_{dir}|y_i)$ sarà stimato dalla media dei j valori simulati:

$$p(\alpha_{dir}|y_i) = (1/j)\sum_{t=1}^{j} p(\alpha_{dir}|\eta_t, s)$$

dove η_t sono i valori campionari di η.
Step 4: Aggiornamento dei parametri complessivi σ^2 *e* π

La varianza σ^2 della componente erratica $\varepsilon_i(t)$ è comune per la componente parametrica e non parametrica e sarà aggiornata tramite la distribuzione congiunta usando tutte le osservazioni.
Per aggiornare π, la probabilità di essere nella componente non parametrica, si estrae dalla distribuzione a posteriori:

$$[\pi|\ldots] \propto Beta(a + \sum_{i=1}^{n_h} I(S_i \neq -1))$$

dove $I(.)$ è la funzione indicatrice.
Per aggiornare σ^2 dobbiamo distinguere i due diversi tipi di contributi dati dalle curve allocate nella componente parametrica e non parametrica.

Per quanto riguarda la componente parametrica, che supponiamo raccolga n_{param} curve, dato che:

$$y_i \sim N_T(X_i\theta_i, \sigma^2 I_T)$$
$$\sigma^{-2} \sim Gamma(a, b)$$

Allora i parametri della distribuzione a posteriori di σ^{-2} che dipendono solo dal contributo delle osservazioni parametriche sono:

$$\begin{aligned} \sigma^{-2} &\sim Gamma(a^*_{param}, b^*_{param}) \\ a^*_{param} &= a + n_{param}/2 \\ b^*_{param} &= b + 0.5(\sum_{i=1}^{n}(y_i - X_i\theta_i)^2 I(S_i = -1) \end{aligned}$$

Per quanto riguarda la componente non parametrica, che supponiamo raccolga n_{non_param} curve, dato che:

$$y_i \sim N_T(\mu, \sigma^2 I_T)$$
$$\sigma^{-2} \sim Gamma(a, b)$$

Allora i parametri complessivi della distribuzione a posteriori di σ^{-2} sono:

$$\sigma^{-2} \sim Gamma(a*, b*)$$
$$a^* = a_{param} + n_{non_param}/2$$
$$b^* = b_{param} + 0.5(\sum_{i=1}^{n_{non_param}} (y_i - \mu)^2$$

3.4 Label switching

Dopo aver impostato il modello bayesiano non parametrico e dopo averlo implementato e ottenuto i risultati in modo iterativo, bisogna riassumere le informazioni ottenute. In pratica bisogna produrre una stima dei parametri del modello e, come in questo caso, precedere alla segmentazione delle osservazioni.
Nell'analisi bayesiana di modelli miscuglio con un numero finito di componenti, si possono avere dei problemi nell stima dei parametri e nella segmentazione delle osservazioni.
In particolare per la stima dei parametri la pratica comune è quella di utilizzare la media della distribuzione a posteriori, e la definizione della distribuzione congiunta a posteriori a partire dalle distribuzioni marginali, ma spesso in questo modo si arriva a dei risultati non buoni.
Questo è dovuto al problema noto come *label switching* che è causato dalla simmetria nella verosimiglianza dei parametri del modello.

Una frequente risposta a questo problema è quello di rimuovere la simmetria utilizzando dei vincoli artificiali di identificabilità.
Nel lavoro di Stephens [2000a] si mostra che spesso questa tecnica fallisce nel risolvere il problema e viene mostrata un classe alternativa di approcci definiti, algoritmi di *relabelling*, che nascono con l'obiettivo di minimizzare la perdita attesa a posteriori sotto una classe di funzioni di perdita.

3.4.1 Introduzione

Il problema del *label switching* sorge quando si affronta il problema della stima dei parametri e della segmentazione in ottica bayesiana utilizzando dei modelli miscuglio.
Il termine *label switching* è stato utilizzato da Redner and Walker [1984] per descrivere l'invarianza della verosimiglianza rinominando (*relabelling*) le componenti del modello miscuglio.
In un contesto bayesiano questa invarianza può condurre ad una distribuzione a posteriori dei parametri fortemente simmetrica e multimodale, rendendo difficile una sua sintesi. In particolare la pratica usuale di sintetizzare la distribuzione congiunta a posteriori marginalizzando la distribuzione e la stima delle quantità di interesse utilizzando la media della distribuzione a posteriori, risulta inappropriata.

3.4.2 Il problema del label switching

Se si indica con $y = y_1, ..., y_n$ il vettore di osservazioni indipendenti da un distribuzione miscuglio con k componenti, dove k è assunto noto e finito:

$$p(y|\pi, \psi, \eta) = \pi_1 f(y; \psi_1, \eta) + \ldots + \pi_n f(y; \psi_n, \eta)$$

dove $\pi = (\pi_1, \ldots, \pi_n)$ sono le proporzioni delle componenti di miscuglio e sono vincolate ad essere non negative e devono sommare ad 1, $\psi = (\psi_1, \ldots, \psi_n)$ sono parametri specifici della componente e η è un parametro comune a tutte le componenti ed f è la funzione di densità.
Si indica con $\theta = (\pi, \psi, \eta)$ il vettore complessivo dei parametri della distribuzione miscuglio.
È talvolta utile assumere che ciascuna osservazione y_i provenga da un'ignota componente z_i della miscuglio, dove $z_1, \ldots, z_n$ sono realizzazioni delle variabile casuali discrete indipendenti e identicamente distribuite $Z_1, \ldots, Z_n$ con funzione di probabilità:

$$Pr(Z_i = j|\theta) = \pi_j (i = 1, \ldots, n; j = 1, \ldots, k)$$

Condizionatamente alle variabili casuali Z, $y_1, \ldots, y_n$ sono osservazioni indipendenti dalle seguenti densità:

$$p(y_i|Z_i = j, \theta) = f(y_i; \psi_j, \eta)(i = 1, \ldots, n)$$

Un approccio bayesiano all'inferenza richiederebbe la specificazione della distribuzione a priori $p(\pi, \psi, \eta)$ per i parametri del modello miscuglio.
L'inferenza sarà quindi basata sulla distribuzione a posteriori $p(\pi, \psi, \eta|y)$, e le quantità di interesse vengono calcolate integrando per i parametri del modello sulla distribuzione a posteriori. Per esempio, la probabilità marginale di classificazione per un'osservazione y_{n+1} è data da:

$$\begin{aligned} Pr(Z_{n+1} = j|y_{n+1}, y) &= \int \frac{\pi_j f(y_{n+1}; \psi_j, \eta)}{\sum_l \pi_l f(y_{n+1}; \psi_l, \eta)} p(\theta|y) d\theta \\ &\sim \int \pi_j f(y_{n+1}; \psi_j, \eta) p(\theta|y) d\theta \end{aligned}$$

Un'accurata approssimazione di questo integrale si ottiene utilizzando i metodi Markov Chain Monte Carlo (MCMC).

3.4.3 Vincoli di identificabilità

Per ciascuna permutazione ν da $1, \ldots, k$ si definisce la corrispondente permutazione del vettore di parametri θ:

$$\nu(\theta) = \nu(\pi, \psi, \eta) = ((\pi_{\nu(1)}, \ldots, \pi_{\nu(k)}), (\psi_{\nu(1)}, \ldots, \psi_{\nu(k)}), \eta).$$

L'origine del problema del *label switching* risiede nel fatto che la verosimiglianza:

$$L(\theta; y) = \prod_{i=1}^{n} \{\pi_1 f(y_i; \psi_1, \eta) + \ldots + \pi_k f(y_i; \psi_k, \eta)\}$$

è la stessa per tutte le permutazioni di θ.
Se nell'analisi bayesiana non abbiamo informazioni a priori per scegliere tra le componenti del modello miscuglio, quindi, se $p(\pi, \psi, \eta)$ è lo stesso per tutte le permutazioni di θ, ci ritroveremo con una distribuzione a posteriori simmetrica. Questa simmetria può portare problemi nel momento in cui si provano ad estrarre informazioni riassuntive sui parametri di interesse da questa distribuzione a posteriori. Per esempio a causa della simmetria le densità predittive delle componenti di scala, date dal secondo integrale precedente, sono le stesse per tutte le componenti, e quindi le probabilità marginale di classificazione date dal primo integrale, sono pari ad $1/k$ per tutte le osservazioni. Queste probabilità di classificazioni sono quindi inutili se l'obiettivo è quello di suddividere le osservazioni in cluster. In modo analogo, la media a posteriori dei parametri specifici delle singole componenti del modello miscuglio sono le stesse per tutte le componenti e sono quindi delle stime scarsamente informative.
Una risposta comune a questo problema è quella di imporre dei vincoli di

identificabilità allo spazio parametrico (per esempio $\pi_1 < \pi_2 < \ldots < \pi_k$ oppure $\psi_1 < \psi_2 < \ldots < \psi_k$) che può essere soddisfatta soltanto per una sola permutazione di θ. Questo rompe la simmetria della distribuzione a priori dei parametri, e quindi della posteriore, e questo dovrebbe risolvere il problema del *label switching*. Putroppo, sulla base di risultati empirici, (si veda Stephens [2000b]) molte scelte dei vincoli di identificabilità si rivelano inutili nel rimuovere il problema della simmetria della distribuzione a posteriori.

3.4.4 Label switching e teoria delle decisioni

E' possibile approcciarsi a questo tipo di problema da un altro punto di vista.
Possiamo vedere la stima dei parametri, la segmentazione delle osservazioni e la sintesi della distribuzione a posteriori come la scelta di un'azione a da un insieme di possibili azioni $\mathcal{A}$.
Seguendo l'approccio della teoria delle decisioni si deve definire una funzione di perdita $\mathcal{L} : \mathcal{A} \times \Theta \to \mathcal{R}$, dove $\mathcal{L}(a;\theta)$ è la perdita che deriva della scelta a quando il vero valore del parametro è θ.
La scelta ottimale della decisione da prendere $\widehat{a}$ è quella che minimizza la perdita attesa a posteriori, cioè il rischio correlato all'azione intrapresa, data da:

$$\mathcal{R}(a) = \mathcal{E}\mathcal{L}(a;\theta|y) \tag{3.6}$$

Bisogna considerare funzioni di perdita invarianti permutazioni dei parametri θ. Per far questo si possono scegliere funzioni di perdita della forma:

$$\mathcal{L}(a;\theta) = min_{\nu}[\mathcal{L}_0(a;\theta)] \tag{3.7}$$

Se $\theta^{(1)}, \ldots, \theta^{(N)}$ sono stati estratti da una *Markov Chain* (dopo un periodo di $burn-in$) con distribuzione stazionaria $p(\theta|y)$, allora una scelta ottimale è quella di approssimare la funzione di rischio $\mathcal{R}(a)$ con il $MonteCarlRisk$, considerando cioè la media dei possibili valori che la funzione di perdita può assumere:

$$\begin{aligned} \widetilde{\mathcal{R}}(a) &= \frac{1}{N}\sum_{t=1}^{N} min_{\nu_t}\left[\mathcal{L}_0\left(a;\nu_t\left(\theta^{(t)}\right)\right)\right] \\ &= min_{\nu_i,\ldots,\nu_N}\left[\frac{1}{N}\sum_{t=1}^{N}\mathcal{L}_0\left(a;\nu_t\left(\theta^{(t)}\right)\right)\right] \end{aligned} \tag{3.8}$$

e si sceglie $\widehat{a}$ che minimizza $\widetilde{\mathcal{R}}(a)$. Per individuare la scelta ottimale si può seguire il seguente algoritmo:

1. scegliere $\widehat{a}$ che minimizza $\sum_{t=1}^{N}\mathcal{L}_0(\widehat{a};\nu_t(\theta^{(t)}))$

2. per ciascun $t = 1, \ldots, n$ scegliere ν_t

La complessità computazionale di questo algoritmo dipende dalla scelta di $\mathcal{L}_0$. Questo algoritmo raggiunge un punto di convergenza, dato che a ciascuna iterazione $\widetilde{\mathcal{R}}$ decresce e c'è un numero finito di possibili valori per le permutazioni $\nu_1, \ldots, \nu_k$. Come molti algoritmi di ottimizzazione la cui ricerca del punto ottimale procede per via iterativa, la soluzione individuata può dipendere dal punto di partenza, e non c'è garanzia nell'individuazione dell'ottimo globale. Per questo motivo è utile eseguire l'algoritmo a partire da diversi punti iniziali e scegliere $\widehat{a}$ che corrisponde al miglior ottimo locale individuato.
Questo algoritmo si rivela particolarmente utile se l'obiettivo principale è la distribuzione a posteriori dei parametri.
In altri contesti si è invece più interessati alla segmentazione delle osservazioni.
Supponiamo di voler utilizzare il modello miscuglio per segmentare le osservazioni in k gruppi e di avere delle informazioni sull'incertezza di questa segmentazione. Un modo naturale per far questo è quello di riportare in una matrice $Q = (q_{ij})$ di dimensioni $n \times k$, dove q_{ij} rappresenta la probabilità che l'osservazione i venga assegnata al cluster j, vincolando le righe di Q a sommare ad 1. Se si interpretano le righe di Q come vettori di probabilità indipendenti, allora Q corrisponde alla distribuzione dei dati in k gruppi.
Se si indica con $P(\theta) = p_{ij}(\theta)$ la matrice delle probabilità di classificazione, dove:

$$p_{ij}(\theta) = Pr(Z_i = j)|y, \theta = \frac{\pi_j f(y_i; \psi_j, \eta)}{\sum_l \pi_l f(y_i; \psi_l, \eta)} \tag{3.9}$$

dove a numeratore abbiamo la densità associata ad y imponendo la configurazione dei parametri ψ del gruppo j, mentre a denominatore abbiamo il fattore di normalizzazione per tale probabilità.
Un modo naturale per misurare la perdita nel riportare Q quando il vero valore del parametro è θ è la distanza di Kullbach-Leibler tra la vera distribuzione nella segmentazione corrispondente a $P(\theta)$ e la distribuzione nella classificazione indicata in Q:

$$\begin{aligned} \mathcal{L}_0 &= \sum_{z_1=1}^{k} \cdots \sum_{z_n=1}^{k} p_{1z_1}(\theta) \ldots p_{nz_n}(\theta) log \left\{ \frac{p_{1z_1}(\theta) \ldots p_{nz_n}(\theta)}{q_{1z_1} \cdots q_{nz_n}} \right\} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}(\theta) log \left\{ \frac{p_{ij}(\theta)}{q_{ij}} \right\}. \end{aligned} \tag{3.10}$$

In pratica l'obiettivo è trovare la configurazione Q il più possibile simile a P sotto il vincolo che le righe di Q devono sommare a 1.

Per questa scelta della funzione di perdita gli step dell'algoritmo iterativa per l'individuazione della soluzione ottima si modificano nel seguente modo:

1. Si inizia con una qualche configurazione iniziale per $\nu_1, \ldots, \nu_N$ e si iterano i seguenti due step fino alla convergenza
2. Si sceglie $\widehat{Q} = (\widehat{q_{ij}})$ che minimizza

$$\sum_{t=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij} \nu_t(\theta^{(t)}) log \left[\frac{p_{ij} \nu_t(\theta^{(t)})}{\widehat{q_{ij}}} \right] \tag{3.11}$$

3. per $t = 1, \ldots, N$ scegliere ν_t che minimizza:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij} \nu_t(\theta^{(t)}) log \left[\frac{p_{ij} \nu_t(\theta^{(t)})}{\widehat{q_{ij}}} \right] \tag{3.12}$$

Alla fine del processo iterativo, quando l'algoritmo di ottimizzazione giunge a convergenza, si può utilizzare la matrice $\widehat{Q}$ per classificare le osservazioni in gruppi scegliendo le variabili di allocazione z_i che massimizzano $\widehat{Q}_{iz_i} (i = 1, \ldots, n)$.
Questi algoritmi per il *relabeling* possono definirsi robusti rispetto alla scelta della funzione di perdita e alla definizione dello spazio delle decisioni, infatti applicando diverse algoritmi per risolvere il problema del *label switching* facendo variare la funzione di perdita e lo spazio delle possibili scelte, si ottengono risultati simili.
Bisogna invece sottolineare che ci sono situazioni in cui si presenta il problema del *label switching* che non può essere risolta con l'algoritmo proposto. Questo succede per esempio nell'analisi bayesiana dei modelli *Hidden Markov* in cui la verosimiglianza è invariante rispetto a permutazioni delle *label* degli stati nascosti.

Capitolo 4

Il problema di marketing e applicazione

I dati a disposizione si riferiscono alla *Customer Base* di una società di telecomunicazioni che per ciascun cliente ha registrato mensilmente il numero di telefonate effettuate, il numero di *sms* e il numero di *mms* inviati da ciascuna *sim card* associata ai clienti.
Inizialmente nell'analisi si volevano considerare solo le informazioni relative al numero di telefonate, ma dopo un'attenta analisi si è deciso di inserire anche le informazioni relative al numero di *sms* e di *mms* inviati.
Questa decisione è stata presa in quanto, di solito, le informazioni di interesse sulle telefonate riguardano le durate e non il numero delle telefonate effettuate. A questo punto, si è deciso di considerare anche il numero di *sms* e *mms* inviati, con l'obiettivo di studiare il generico andamento dell'utilizzo dell'apparecchio telefonico.
L'orizzonte temporale dei dati a disposizione inizia a Novembre 2004 e termina ad Aprile 2006.
L'operazione preliminare da fare prima di intraprendere un'analisi di *data mining* è quella di ripulire il database da osservazioni anomale.
La base dati disponibile comprende 32.524 osservazioni. Dopo aver eliminato i record che presentavano dei livelli anomali nell'utilizzo del telefono, per esempio le serie a domanda intermittente o, che in casi limite, presentano solo un punto diverso da zero, si arriva ad un *dataset* finale composto da 20.809 serie storiche. Da questo *dataset*, per la stima del modello bayesiano, è stato infine estratto un campione casuale di 1.000 osservazioni.

4.1 Modello miscuglio bayesiano

Le serie storiche di interesse sono costituite da 18 osservazioni mensili e possiamo indicarle con il vettore:

$$y_i = (y_{i,1}, ..., y_{i,18}) \tag{4.1}$$

Applicando il modello descritto nel Paragrafo 3.3 dopo aver effettuato 4.016 iterazioni con un periodi di *burn in* di 500 iterazioni, si ottengono i seguenti risultati.
Si osserva dai *trace-plot* di alcun parametri in Figura 4.1 che il processo iterativo sembra essere arrivato a convergenza.
Alla fine del processo iterativo, e dopo aver eliminato le prime iterazioni re-

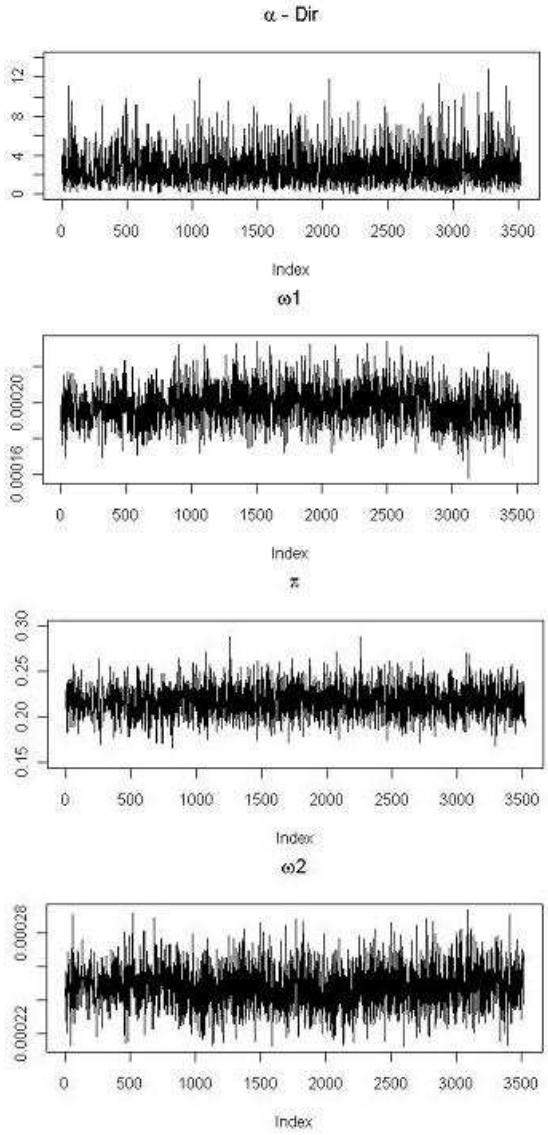

Fig. 4.1: *Trace plot* dei parametri del modello bayesiano (α_{Dir},π,ω_1,ω_2).

lative al periodo di *burn in* abbiamo a disposizione un insieme di circa 3500 configurazioni dell'allocazione delle *sim card* nella componente parametrica

e non parametrica con i relativi parametri stimati per procedere all'allocazione.
Per avere un'idea dell'andamento che caratterizza le serie storiche assegnate alla componente parametrica, si riporta in Figura 4.2 il grafico con i livelli del traffico telefonico di una *sim card* assegnata alla componente parametrica. In questo figura, la serie storica relativa alla *sim card* della *customer base* assegnata alla componente parametrica,è rappresentata tramite i punti, e si può vedere come la stima della distribuzione ottenuta applicando il modello parametrico, indicata con la linea continua che interpola i punti osservati, sia una buona stima sia in termini di bande di variabilità che risultano sufficientemente strette, sia in termini di forma. Si nota la differenza con la linea tratteggiata in basso al grafico che rappresenta la stima che si sarebbe ottenuta considerando la media di tutte le *sim card* analizzate.

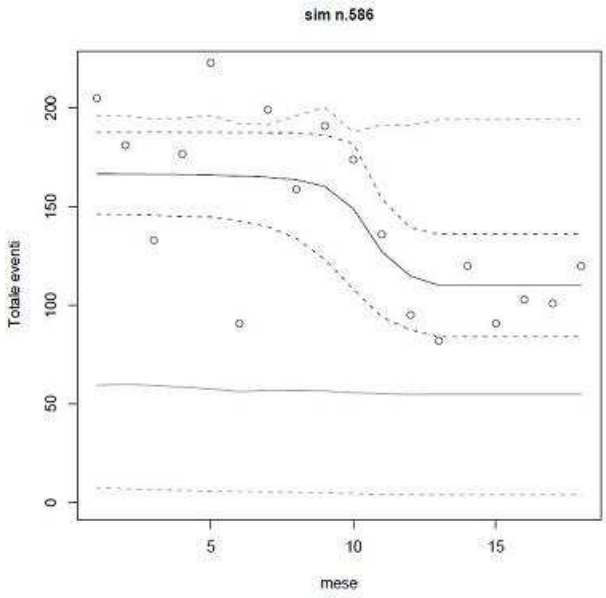

Fig. 4.2: Andamento del traffico telefonico di una *sim card*

Analizzando la media e la deviazione standard dei parametri principali del modello parametrico, in Figura 4.3 e in Figura 4.4, possiamo vedere che nel gruppo dei clienti che presentano un andamento decrescente, i livelli iniziali del traffico telefonico si distribuiscono intorno ad un valore di 60 considerando il numero di telefonate effettuati e di messaggi inviati. La distribuzione è poco concentrata intorno a questo valo in quanto si osserva una deviazione standard in media di 70 unità.
La stessa variabilità elevata si osserva nel parametro $\theta_{i,2}$ che rappresenta il coefficiente angolare delle spezzata che congiunge il *plateau* iniziale a quello finale dei livelli del traffico telefonico. Questo parametro assume valori negativi, parliamo quindi di 'Decremento totale', la cui distribuzione presenta

come valore medio -5.
Più interessanti risultano i parametri k_i e r_i della componente parametrica

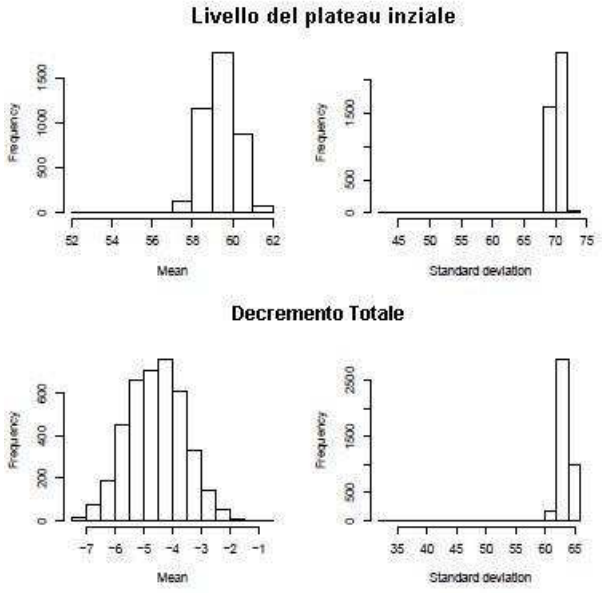

Fig. 4.3: Media e deviazione standard del livello iniziale del traffico telefonico e del tasso di decremento totale.

del modello in Figura 4.4. Si ricorda che questi due parametri si riferiscono rispettivamente all'istante temporale (in questo caso il mese) in cui il traffico telefonico dell'utente comincia a diminuire e al periodo di tempo in mesi impiegato per raggiungere il livello finale caratterizzato da consumi ridotti. Questi due parametri presentano una distribuzione sufficientemente concentrata intorno alla loro media. Si può quindi affermare che il momento in cui il livello del traffico telefonico comincia a diminuire si colloca interno al sesto mese dell'orizzonte di studio e che questo periodo di calo durerà per circa 2 o 3 mesi.
Si potrebbe quindi anticipare questa tendenza analizzando i fattori esogeni che si verificano in questo periodo di tempo e che sono correlati con l'utilizzo dell'apparecchio telefonico.
In Figura 4.5 abbiamo la distribuzione delle percentuali di allocazione nelle componente non parametrica di ciascuna serie storica. Da questo output si nota che c'è una buona quota di *sim card* che non rientra mai nella componente non parametrica, inoltre si può decidere di eliminare dalla componente non parametrica alcune osservazioni che sono state allocate in questa componente con bassa frequenza. Considerando la distribuzione di queste percentuali di allocazione si fissa come *cut-off* di uscita dalla componente non parametrica l'80%.

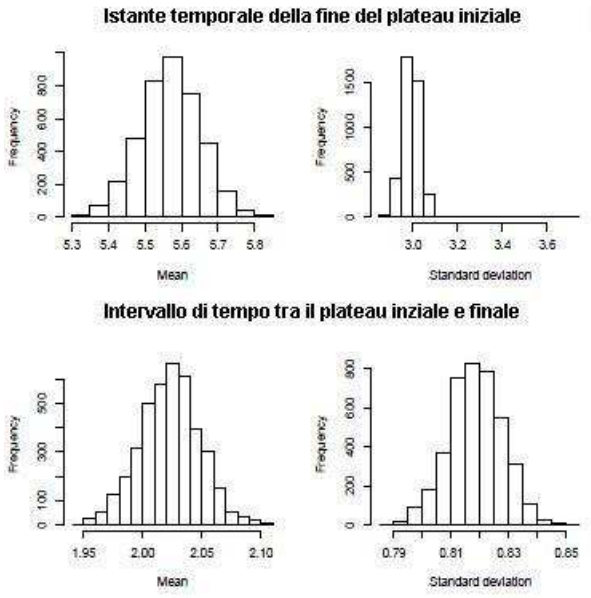

Fig. 4.4: Media e deviazione standard dell'istante in cui il livello del traffico telefonico comincia a diminuire e durata di questa fase di decremento.

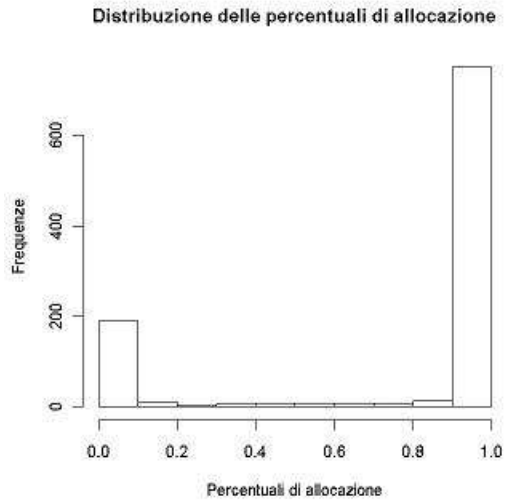

Fig. 4.5: Distribuzione delle percentuali di allocazione delle *sim card* nella componente non parametrica.

In questo modo focalizziamo l'attenzione su quelle serie storiche che con elevata probabilità appartengono alla componente non parametrica, e tra queste cerchiamo di verificare se esistono dei *pattern* più frequenti degli altri, e che più si differenziano dal resto delle serie storiche.
Per quanto riguarda le singole componenti non parametriche, in ogni iterazione ne abbiamo un numero diverso, di cui se ne riporta la distribuzione in Figura 4.6. Vediamo che la frequenza del numero di componenti non parametriche si distribuisce intorno a 12 con una coda piuttosto pesata verso valori più elevati. Il valore medio del numero di gruppi individuato, pesato per il numero di volte in cui questo valore compare nelle iterazioni è di circa 15.

Tra tutte le componenti non parametriche create, solo alcune di queste

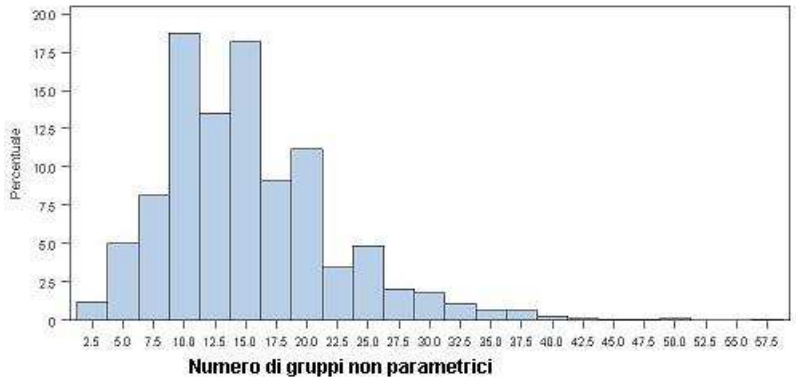

Fig. 4.6: Distribuzione del numero di componenti non parametriche individuate in ciascuna iterazione.

raccolgono una quota elevata di osservazioni, altre componenti sono invece marginali, al loro interno cadono poche serie storiche. Anche in questo caso si potrebbe focalizzare l'attenzione sulle componenti non parametriche più presenti in ciascuna iterazione e vedere in che modo le singole unità statistiche vengono allocate in ciascuna di queste.
In media in ciascun iterazione vengono creati 15 gruppi non parametrici, ma al massimo solo tre di questi gruppi riescono a raggruppare un numero di *sim card* sufficientemente elevato.
In Figura si mostra la distribuzione del numero di *sim card* allocate nelle singole componenti non parametriche in tutte le iterazioni effettuate. Ci sono molte componenti parametriche marginali che comprendono solo poche osservazioni, mentre si osserva che ci sono delle concentrazioni elevate intorno a 120 e 480. Fissando quindi a 100 la numerosità minima delle *sim card* cadute nelle singole componenti non parametriche che si vogliono considerare, in ciascuna iterazione si riescono ad individuare al massimo quattro di queste componenti non parametriche.

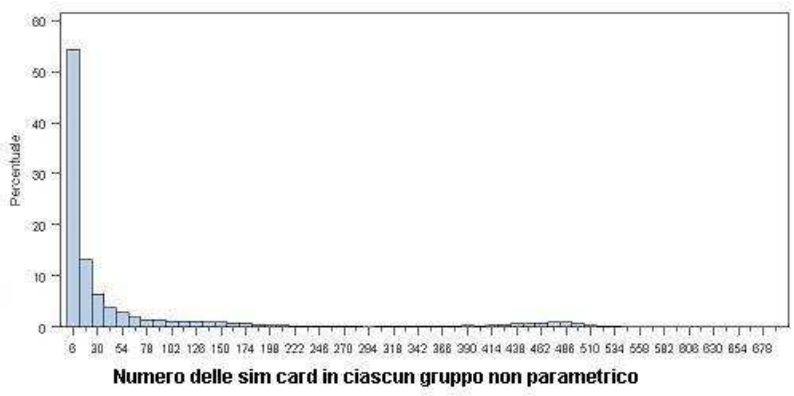

Fig. 4.7: Distribuzione del numero di sim card allocate nelle singole componenti parametriche.

Tab 1. : Numero di gruppi non parametrici che contengono almeno 100 *sim card*.

Numero di gruppi	Frequenza	Percentuale	Frequenza Cumulata
1	3473	48.93	3473
2	2950	41.56	6423
3	642	9.04	7065
4	33	0.46	7098

Dalla Figura 4.8, in cui si vedono le medie delle componenti non parametriche appartenenti a ciascuna delle due tipologie non parametriche, si riesce a caratterizzare il profili dell'andamento di queste due tipologie di componenti non parametriche. Entrambe le componenti non parametriche presentano un avvallamento in corrispondenza del mese di Aprile e:

- la Tipologia 1 dopo Aprile presenta una fase di calo costante
- la Tipologia 2 dopo Aprile presenta una fase di ripresa seguita da un calo.

Alla fine dell'analisi abbiamo quindi individuato un raggruppamento delle unità in tre gruppi, uno parametrico e due non parametrici.

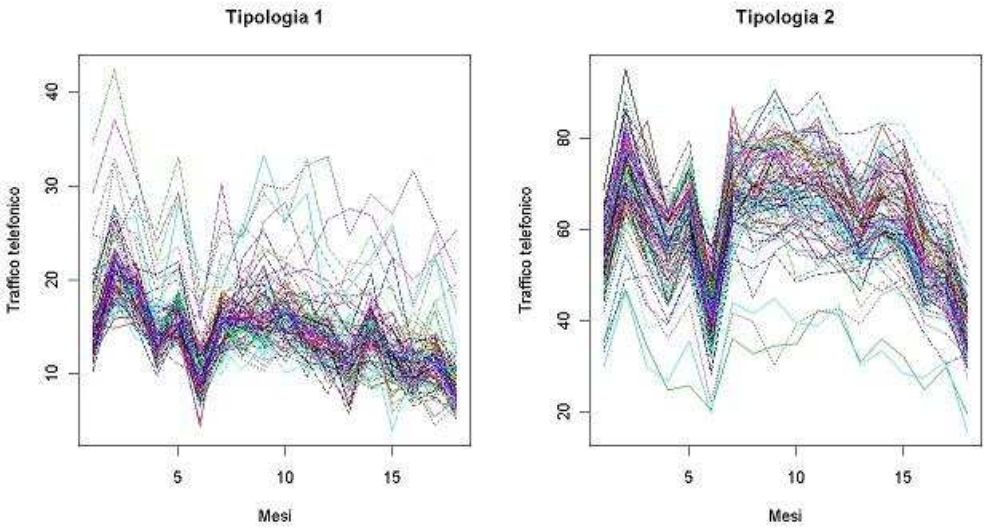

Fig. 4.8: Distribuzione delle medie delle due principali tipologie di componenti non parametriche.

Capitolo 5

Conclusione

L'obiettivo della tesi di individuare tramite un modello miscuglio non parametrico bayesiano gli andamenti tipici del traffico telefonico degli utenti di una compagni di telefonia, è stato raggiunto con risultati interessanti.
Grazie alla flessibilità del modello miscuglio bayesiano utilizzato per affrontare il problema di marketing di segmentazione delle clientela, si sono individuati due andamenti tipici nell'utilizzo del telefono cellulare, in aggiunta all'andamento definito della componente parametrica.
Riguardo alla componente parametrica siamo riusciti a ottenere informazioni più dettagliate riguardo alla forma specifica di questo andamento.
In particolare dell'andamento decrescente del traffico telefonico, di maggior interesse per l'azienda, si è individuato per ciascun cliente l'istante temporale in cui questo calo si manifesta e il periodo di tempo in cui i livelli del traffico continuano a diminuire fino ad un livello costante finale.
Tutte queste informazioni oltre ad essere di per sé interessanti per una conoscenza approfondita delle modalità di fruizione dei servizi di telefonia, possono essere utili per prevenire eventuali eventi di *churn*, andando ad intervenire su quei fattori che influenzano il comportamento dei clienti.
Imponendo la struttura lineare a tratti della componente parametrica e permettendo ai parametri di aggiornarsi sulla base dei dati osservati, è stato possibile modellare contemporaneamente gli andamenti caratterizzati da cali più o meno veloci con la presenza eventuale di *plateau*, di ampiezza e livello variabile, all'inizio e alla fine del periodo di osservazione.
Inoltre, l'approccio utilizzato può essere reiterato più volte parametrizzando le nuove componenti non parametriche individuate utilizzandole quindi come informazioni a priori del processo.

Bibliografia

J. Bigelow and David B.Dunson. Bayesian adaptive regression splines for hierarchcal data. *Biometrics*, 63(3):724–732, September 2007.

Luisa Bisaglia, Margherita Gerolimetto, and Bruno Scarpa. Statistical analysis for customer profiling. *Proceedings of the 2007 intermediate conference*, June, 2007.

Maela Bonetto. Prevedere il churn: un approccio longitudinale. *Tesi di laurea Triennale*, 2007.

Giacomo Cassol. Il valore del cliente: analisi di classificazione con dati longitudinali. *Tesi di laurea Specialistica*, 2009.

Selina Chu, Eamonn Keogh, David Hart, and Michael Pazzani. Iterative Deepening Dynamic Time Warping for Time Series. *Proc* 2^{nd} *SIAM International Conference on Data Mining*, 2002.

Marcella Corduas and Domenico Piccolo. Time series clustering and classification by the autoregressive metric. *Comput. Stat. Data Anal.*, 52(4), 2008.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90 (430), June 1995.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 2(1):209–230, 1973.

Sergio M. Focardi. Clustering delle serie storiche economiche:Applicazioni e questioni computazionali. *The Intertek Group*, Ottobre 2001.

Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31, 2009.

Amara L. Graps. An Introduction to Wavelets. *IEEE Computational Sciences and Engineering*, 2(2):50–61, Summer 1995.

Nicholas A. Heard, Christopher C. Holmes, and David A. Stephens. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, 101:18–29, 2006.

Flip Korn, H. V. Jagadish, and Christos Faloutsos. Efficiently supporting ad hoc queries in large datasets of time sequences. *Proceedings of SIGMOD '97*, 1997.

James O. Ramsay and Bernard W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, 2002.

Shubhankar Ray and Bani Mallick. Functional clustering by bayesian wavelet methods. *Journal Of The Royal Statistical Society Series B*, 68(2):305–332, 2006.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Rev.*, 26:195–239, 1984.

Abel Rodriguez, David B. Dunson, and Alan E. Gelfand. Bayesian nonparametric functional data through density estoimation. *Biometrika*, 96(1): 149–162, 2009.

Bruno Scarpa and David B. Dunson. Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors. *Biometrics*, 65(3):772–780, September 2009.

Matthew Stephens. Dealing with label switching in mixture models. *J.R. Statist. Soc. B*, 62:795–809, 2000a.

Matthew Stephens. Bayesian analysis of mixture models with an unknow number of components. an alternative to reversible jump methods. *Annals of Statistics*, 28, 2000b.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:1566–1581, December 2006.

Wes Thompson and Ori Rosen. A bayesian model for sparse functional data. *Biometrics*, 64(1), March 2007.

Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal data*. Springer, 2000.