UNIVERSITÀ DEGLI STUDI DI PADOVA

CORSO DI LAUREA MAGISTRALE IN
SCIENZE STATISTICHE

TESI DI LAUREA

Catene di Markov nascoste bayesiane non parametriche con stati persistenti. La previsione dei prezzi dei voli aerei.

RELATORE: Bruno Scarpa
DIPARTIMENTO DI SCIENZE STATISTICHE

LAUREANDO: Gianluca Barbierato

ANNO ACCADEMICO 2011/2012

Indice

Introduzione

Da circa vent'anni, il progresso tecnologico informatico ha cambiato profondamente il rapporto tra le aziende e l'analisi dei dati.
Grazie alla diffusione del web e alla sempre maggior potenza di calcolo a disposizione, molte imprese possono estrarre valore dall'analisi dei comportamenti del consumatore.
Queste innovazioni, non solo hanno cambiato il modo di utilizzare i dati delle aziende esistenti ma hanno anche creato nuovi tipi di imprese che non potevano esistere solo pochi anni fa.
E' il caso degli *on line store*, cioè delle aziende che vendono prodotti tramite un sito web. Dall'apertura di Amazon, un numero sempre maggiore di aziende ha iniziato a utilizzare questo canale di vendita che permette da una parte di ridurre i costi di gestione e dall'altra di raccogliere facilmente un'enorme quantità di dati con poco sforzo.
Anche l'azienda di cui si parla in questa tesi appartiene a questo gruppo. Bravofly, infatti, è un'azienda che gestisce alcuni siti web, dai quali è possibile ricercare e acquistare viaggi aerei.
L'analisi di cui l'azienda ha bisogno, però, non riguarda il comportamento dei consumatori, o almeno non direttamente. Essa infatti, si pone il problema di cercare di prevedere l'andamento nel tempo dei prezzi dei voli applicati dalle diverse compagnie aeree. Come è noto, chi acquista un volo pochi giorni prima della partenza spesso incorre in un prezzo diverso rispetto a chi l'ha acquistato con qualche giorno, o addirittura qualche mese in più di anticipo. Per determinare il prezzo dei voli, le compagnie aeree si servono di

molte informazioni, tra le quali vi sono, ad esempio, il numero di posti già venduti e il numero di posti venduti in passato per voli con caratteristiche simili.
Bravofly, però, non può disporre di questo tipo di informazioni. Tutto quello che ha a disposizione è un insieme di andamenti osservati nel passato per il prezzo di voli.
Dal punto di vista statistico, il problema consiste perciò nella modellazione di dati funzionali. Questo problema ricorre in molti ambiti. Dunson (2010), ad esempio, ha studiato dati funzionali relativi all'andamento nel tempo dei livelli di progesterone nei primi mesi di gravidanza, mentre Wang, Jank e Shmueli (2008) si sono occupati della dinamica dei prezzi delle aste *on line*.
In questa tesi verrà utilizzato un approccio Bayesiano non parametrico per l'analisi dei dati. Questo tipo di approccio permette di superare alcuni limiti della statistica classica e ha portato a risultati molto interessanti riguardanti, ad esempio, le interrogazioni ai motori di ricerca (Cowans, 2004) e il riconoscimento vocale (Goldwater, Griffiths e Johnson, 2006).
L'esposizione è strutturata come segue. Nel primo capitolo si espone meglio il problema che si vuole affrontare e si presenteranno i dati forniti da Bravofly con le relative modifiche effettuate sulla loro struttura.
Nel secondo capitolo verranno introdotti la statistica Bayesiana non parametrica e il processo di Dirichlet e verrà presentato un primo modello basato sul processo di Dirichlet funzionale (MacEachern, 1994).
Nel terzo, verrà introdotta la struttura probabilistica delle catene di Markov nascoste (Rabiner, 1989) e si mostrerà come applicare il processo di Dirichlet gerarchico (Teh, Jordan, Beal e Blei, 2006) per modellare catene di Markov con un numero non predeterminato di stati. Verrà poi introdotta una variazione per la modellazione di catene con stati persistenti per ottenere così il modello *sticky* HDP-HMM (*hierarchical Dirichlet process hidden Markov model*; Fox, Sudderth, Jordan e Willsky, 2008).
Infine, nel quarto capitolo verrano presentati i risultati ottenuti dall'applicazione di entrambi i modelli. In particolare, si mostrerà come è possibile ot-

tenere un raggruppamento delle funzioni di prezzo osservate tramite il modello basato sul processo di Dirichlet funzionale (FDP) mentre si cercherà di capire se esistono alcuni momenti particolari in cui è lecito attendersi una sensibile variazione di prezzo usando il modello presentato nel terzo capitolo.

Capitolo 1

Il problema di business e i dati

Per le aziende operanti in settori molto dinamici come quello dei viaggi aerei, è molto importante avere informazioni su quello che sta accadendo nel mercato il più velocemente possibile.
Nel primo paragrafo di questo capitolo verrà esposto il problema di business che si trova ad affrontare Bravofly. Esso riguarda principalmente la previsione dell'andamento del prezzo nel tempo di un viaggio aereo e, più in generale, la possibilità di avere informazioni sulla forma di queste funzioni.
Il prezzo di prodotti di questo tipo varia nel tempo soprattutto in base a fattori difficilmente controllabili dalle aziende come Bravofly. La maggior parte dei modelli fin qui utilizzati in letteratura opera una semplificazione che prevede di studiare il problema in soli due istanti temporali. Questa semplificazione non è accettabile per gli scopi che si prefigge l'impresa dato che essa è interessata a prevedere giorno per giorno il prezzo del volo.
Nei due paragrafi successivi si presenteranno la struttura dei dati forniti e le elaborazioni condotte su di essi per poterli usare nei modelli descritti nei capitoli successivi.
Per la lettura e il trattamento dei dati si è usato il software SAS System 9.2, presente nelle aule ASID del Dipartimento di Scienze Statistiche dell'Università di Padova.

1.1 L'azienda e il problema di business

Il gruppo Bravofly è uno dei leader nel settore della vendita di viaggi *on line*. Il suo *core business* è dato dalla vendita di voli aerei ma dai vari siti web gestiti dal gruppo l'utente può comprare anche viaggi in nave o traghetto, interi pacchetti vacanze, prenotare hotel, ristoranti e noleggiare automobili.
Il successo dell'azienda è da individuarsi principalmente nel suo motore di ricerca. Ogni volta che un utente compone una richiesta, il sito deve cercare tutti i voli disponibili tra tutte le compagnie aeree, per fornire all'utente una vasta gamma di soluzioni ordinate per aderenza alle preferenze descritte e per prezzo.
In particolare, la piattaforma di ricerca Bravofly permette ai consumatori di cercare, con una sola richiesta, soluzioni *low cost* e tradizionali, e confrontare le proposte trovate non solamente in base al prezzo ma anche in base a caratteristiche di comodità come orario di partenza e arrivo, numero di scali, tempo di volo, terminal...
Come si può facilmente intuire, Bravofly è un'azienda per cui le capacità tecnologiche e informatiche sono di primaria importanza. Ogni volta che un utente richiede delle informazioni per un volo, infatti, il motore di ricerca deve sondare tutte le possibilità indagando, di fatto, oltre alle informazioni conservate dalle precedenti ricerche, tutti i server delle compagnie aeree che forniscono quel determinato tragitto. Deve poi verificare la disponibilità di posti, e deve infine riportare il prezzo e tutte le condizioni di viaggio per ognuna di queste possibilità. Il tutto deve essere fatto in un tempo ragionevole (non più di un minuto) in modo da non permettere all'utente di stancarsi e uscire dal sito.
Può capitare, d'altra parte, che per problemi legati alla temporanea inaccessibilità dei server delle compagnie aeree, non sia possibile reperire in quel momento informazioni relative a una determinata soluzione nota da ricerche precedenti (e nemmeno verificarne la disponibilità).
In questi casi, evitare di fornire all'utente la soluzione in questione non è economicamente auspicabile. E' preferibile invece, cercare di sfruttare le in-

formazioni in possesso (sia quelle attuali, sia quelle 'storiche') per proporre al consumatore un prezzo quanto più possibile vicino alla realtà.
Generalizzando, ciò che serve all'azienda è la possibilità di prevedere l'andamento delle curve di prezzo dei voli, basandosi sulle caratteristiche del viaggio, come compagnia aerea e tragitto, e sullo storico delle richieste per quel volo fino a quel momento.
In economia, esiste una vasta letteratura riguardante la gestione nel tempo dei prezzi di beni quali i viaggi aerei (Courty e Li, 2000; Nocke e Peitz, 2008; Watanabe e Muller, 2010). Questo tipo di prodotto rientra in una famiglia che comprende, ad esempio, anche i biglietti per assistere ad eventi sportivi o a spettacoli teatrali e le camere disponibili in un albergo in un dato periodo. Ci sono due principali caratteristiche che accomunano tutti questi prodotti. La prima consiste nel fatto che il loro consumo avviene in un momento predeterminato e il cliente può decidere se comprare il prodotto con largo anticipo o aspettare fino all'ultimo giorno utile. La seconda caratteristica, invece riguarda la capacità produttiva: essa è limitata e non può essere cambiata facilmente (si pensi ad esempio, a quanto può costare, non solo in termini economici, ad una compagnia aerea comprare un nuovo aeroplano).
Quando un'azienda vende servizi di questo tipo, uno dei suoi principali problemi consiste proprio nella determinazione del prezzo. Un prezzo troppo alto, potrebbe far calare troppo la domanda, lasciando parte della capacità produttiva inutilizzata (aereo che parte mezzo vuoto o stadio che non viene riempito). Un prezzo più basso, invece, potrebbe saturare la capacità produttiva ma potrebbe causare guadagni più bassi (è più conveniente vendere 80 biglietti a 2 £ che 100 a 1.5 £).
D'altra parte, problemi del genere riguardano tutti i prodotti e i servizi che un'azienda può vendere che sono soggetti a domanda variabile. Quello che cambia nel caso dei viaggi in aereo o dei biglietti per eventi di intrattenimento è che in questo caso c'è la possibilità di effettuare discriminazioni di prezzo.
Le discriminazioni di prezzo sono di tre tipi. La prima, consiste nel differente livello di servizio offerto (*economy* o *business class*, curva o tribuna, platea o

galleria). La seconda è legata invece alle caratteristiche dell'utente, quali età, sesso e livello di fedeltà. La terza, infine, si traduce in un prezzo che varia in base al tempo che manca al consumo del prodotto in questione, ed è il tipo di discriminazione che si cercherà di studiare in questa tesi.
L'ottimizzazione del volume d'affari derivante dalla vendita di questo tipo di prodotti tramite la gestione del prezzo di vendita, è detta gestione dei ricavi, ma viene usato ormai diffusamente il termine inglese *revenue management* (Netessine e Shumsky, 2002). Esso nacque tra la fine degli anni '70 e l'inizio degli anni '80, presso alcune compagnie aeree americane che lo usarono inizialmente per fronteggiare la concorrenza delle allora nascenti compagnie *low cost*.
Ritornando alla gestione del prezzo nel tempo, in essa si possono distinguere due strategie di base. In alcuni casi, il cliente sa con certezza che consumerà il prodotto in questione, per cui è portato ad effettuare l'acquisto in anticipo, per evitare di correre il rischio che il prodotto si esaurisca. In questi casi, la strategia ottimale per il venditore consiste nel far pagare di più ai clienti che comprano prima, facendo scendere il prezzo man mano che il tempo di consumo si avvicina.
Altre volte invece, si è ragionevolmente sicuri che il prodotto non si esaurirà e il fattore di incertezza riguarda l'effettiva possibilità di poter usufruire del prodotto. In questi casi, il cliente è portato ad effettuare l'acquisto il più tardi possibile e pertanto, la strategia di vendita che ne deriva consiste nell'erogazione di sconti per l'acquisto anticipato.
Ovviamente, nei casi reali non si verificano quasi mai situazioni così semplici, ma si assiste a una implementazione di entrambe le strategie. Ciò avviene anche per i voli aerei: nonostante sia ormai noto che, sostanzialmente il prezzo di un volo aereo aumenta man mano che si avvicina il giorno del viaggio, raramente si assiste a una curva di prezzo monotona.
Come si è detto, esistono molti lavori in letteratura che trattano questo problema, anche in relazione a diversi aspetti che si possono realizzare durante la vendita. Ad esempio Muller e Watanabe (2010) hanno impostato un mo-

dello a due periodi con un venditore monopolista in cui i clienti sono certi del valore che ha per loro la merce solamente nel secondo periodo. Tra le ipotesi hanno considerato inoltre che la capacità fosse esogena, che il venditore dovesse scegliere in anticipo il prezzo del bene per entrambi i periodi e che i clienti non potessero rivendere il prodotto acquistato. Essi hanno dimostrato che nessuna delle strategie semplici descritte è la migliore in assoluto ma la scelta dipende dai parametri del modello e quindi dalle condizioni del mercato cui si fa riferimento. In seguito, hanno verificato che cosa comportasse il cambiamento di alcune delle ipotesi precedenti.
Il lavoro di Watanabe e Muller è tra i più aggiornati in questo ambito, tuttavia, resta comunque un problema di fondo nei modelli da loro analizzati. Essi, infatti, discretizzano il tempo considerando solamente due diversi periodi. Questa è una semplificazione che mal si sposa con le dinamiche di prezzo riscontrate empiricamente, soprattutto nel campo dei viaggi aerei, dove l'acquisto può essere effettuato anche molti mesi prima del consumo.
In questa tesi, si cercherà quindi di considerare un discretizzazione più fine del tempo per cercare di prevedere le complesse dinamiche di prezzo che caratterizzano i dati.
E' bene notare, a questo punto che l'azienda non possiede tutte le informazioni di cui può disporre la compagnia aerea, ad esempio essa non può conoscere con esattezza quanti posti sono stati venduti per il volo in considerazione per due motivi. Il primo è dato dal fatto che a ogni richiesta non necessariamente corrisponde un acquisto. Inoltre, va considerato che Bravofly non è l'unico canale di vendita.
In ogni caso, la previsione dell'aggiustamento dei prezzi dei voli nel tempo da parte delle compagnie aeree è di primaria importanza. Un esempio tra le possibili applicazioni è dato dall'opportunità di offrire all'utente come servizio a pagamento la facoltà di riservarsi il posto per un certo periodo di tempo (ad esempio 24 ore) senza essere poi vincolato all'acquisto. Infine, una buona previsione potrebbe anche velocizzare la ricerca riducendola solamente alla soluzioni per cui le informazioni in possesso sono considerate

ormai obsolete.

1.2 Modalità di raccolta dati

Ogni volta che un utente si connette a uno dei siti web gestiti da Bravofly può richiedere, con pochi e semplici passi, informazioni riguardo a un viaggio aereo impostando gli aeroporti, o le città di partenza e arrivo, e il periodo di interesse. Quando ciò accade, il sito richiede ai server delle compagnie aeree che offrono voli sulla tratta scelta, il prezzo dei singoli voli. In questo modo è stato creato il database originale, in cui in ogni riga sono contenute tutte le informazioni relative a una *hop* al momento della richiesta.
Una *hop* consiste in una singola tratta aerea semplice. A titolo esemplificativo, un viaggio di sola andata da Roma a Mosca può essere costituito da una sola *hop* se non ci sono scali o da due *hop*, ad esempio Roma-Monaco e Monaco-Mosca. In questo secondo caso, il viaggio (o meglio, il preventivo di viaggio) occupa due record.
Per quanto riguarda i viaggi andata e ritorno, va considerata anche la suddivisione in *leg*, che sostanzialmente dividono le *hop* relative a un viaggio in *hop* di andata e ritorno. Nei viaggi di sola andata, invece, si è in presenza di un'unica *leg*.
Dai codici identificativi di ogni *hop* si può ricostruire la composizione del viaggio richiesto. Va qui notato che non necessariamente il tragitto di andata è uguale a quello di ritorno, sia per quanto riguarda il numero di *hop*, sia per quanto riguarda gli aeroporti di partenza e arrivo. Nella tabella 1.1 sono riportati alcuni esempi.
Tutte queste considerazioni relative al contesto di viaggio di ogni tratta aerea considerata non sono da ritenersi rilevanti soltanto da un punto di vista informatico, ma anche e soprattutto da un punto di vista economico. Va sottolineato infatti, che il prodotto commerciale venduto da una compagnia aerea o da un'agenzia viaggi non consiste in una semplice "somma di *hop*". Al contrario, il prezzo di ogni singola tratta varia a seconda di molti fattori, uno dei

ID FLIGHT	ID LEG	ID HOP	...	LEG AMOUNT
1	xxxxxx-1-1	1	...	88.00
1	yyyyyy-1-1	1	...	25.40
1	yyyyyy-1-2	2	...	30.99
1	zzzzzz-1-1	1	...	215.00
1	zzzzzz-1-1	2	...	215.00
1	kkkkkk-1-1	1	...	40.00
1	kkkkkk-1-1	2	...	40.00
1	kkkkkk-1-2	1	...	45.00
1	kkkkkk-1-2	2	...	45.00

Tabella 1.1: *Versione iniziale del database. La prima riga rappresenta un viaggio di sola andata comprendente una tratta unica. La seconda e la terza riga rappresentano un viaggio con una hop per l'andata e una per il ritorno. La terza e la quarta riga mostrano un viaggio di sola andata consistente in due hop. Le ultime quattro righe formano un viaggio con due hop per l'andata e due per il ritorno.*

quali è la sua collocazione all'interno dell'organizzazione del viaggio completo acquistato. Si noti infatti, che il prezzo riportato nella colonna LEG AMOUNT è, come indica l'intestazione, il prezzo relativo alla *leg* completa e che esso viene riportato uguale in più righe quando una *leg* è composta da più *hop*.
Per questo motivo, nelle analisi che verranno presentate sono stati considerati solamente preventivi relativi a:

- voli di sola andata diretti (cioè con una sola *hop*).
- voli andata e ritorno costituiti da due sole *hop* (una per l'andata e una per il ritorno) identiche, considerando in modo indipendente il volo di andata e quello di ritorno.

Le analisi delle curve di prezzo relative a *hop* inserite in viaggi più complessi viene lasciata a lavori successivi. Esse necessiteranno, oltre ai risultati delle

analisi di questa tesi, di altre considerazioni sia di natura economica che statistica.
Per ognuno di questi record presenti nella tabella originaria, oltre alle variabili già mostrate nella tabella 1.1, erano presenti le seguenti variabili:

1. data e ora in cui è stata effettuata la richiesta
2. aeroporto di partenza (codice di tre lettere)
3. data e ora di partenza
4. aeroporto di arrivo
5. data e ora di arrivo
6. codice della compagnia aerea
7. numero del volo (codice internazionale)

Infine, Bravofly ha fornito anche una tabella in cui sono presenti le principali caratteristiche di tutti gli aeroporti presenti nel database.

1.3 Sistemazione dati

La maggior parte delle informazioni contenute nel database originale è presentata in modo ridondante.
Ad esempio, tutte le caratteristiche statiche di ogni viaggio aereo (*hop*) sono riproposte ogni volta che viene presentato un nuovo preventivo per esso. Per di più, spesso un utente compone la stessa richiesta due volte nell'arco di pochi secondi, generando così nel database due record completamente uguali.
Un altro problema è dato dalla mancanza di un vero e proprio identificatore per le *hop*. La variabile ID presente è collegata ai viaggi completi. Per analizzare l'andamento del prezzo di ogni volo, è desiderabile invece che a ogni

riga corrisponda una sorta di serie storica relativa alla *hop* in questione. In quest'ottica, si sono effettuate alcune elaborazioni dei dati forniti dall'azienda.
Per prima cosa, si è fatto in modo che ad ogni preventivo richiesto corrispondesse esattamente una sola riga del database. Per far ciò, è stata usata la prima parte della variabile ID LEG come identificativo delle nuove unità statistiche, mentre, per mantenere tutte le informazioni relative alle diverse tappe del viaggio sono state create nuove variabili (eventuamente lasciate vuote) relative all'aeroporto e all'orario di partenza e arrivo relative a ogni singola *hop*. A questo punto, si sono potuti eliminare dal database i viaggi che non corrispondevano alle caratteristiche descritte a pagina 7.

DEP H1	**DEP DATE H1**	**ARR H1**	**ARR DATE H1**	**DEP H2**	...
AGA	GG-MM-AA HH	ORY	GG-MM-AA HH	ORY	
BVS	GG-MM-AA HH	NAP	GG-MM-AA HH		

Tabella 1.2: *Versione del database in cui a ogni preventivo corrisponde un solo record. Nella seconda riga il viaggio è costituito di una sola hop e pertanto la variabile DEP H2 non ha valore.*

Dopo queste operazioni, si sono create tre tabelle separate: una con i voli di sola andata, una con quelli di andata per i quali è stato prenotato il ritorno e uno per i voli di ritorno. Le successive elaborazioni si sono ripetute per ognuna di queste tre tabelle. Dopo l'eliminazione di alcune variabili, esse contenevano: REQUEST DATE, DEP AIRPORT, ARR AIRPORT, DEP DATE, ID AIRLINE, LEG AMOUNT.
Per poter costruire la serie storica dei prezzi di ogni singolo volo, era necessario un nuovo identificatore che individuasse stavolta, non il preventivo, ma il volo per cui esso è stato richiesto. A questo scopo, è stata creata la variabile NEW, contenente una stringa formata dalla concatenazione dei valori delle variabili DEP AIRPORT, ARR AIRPORT, ID AIRLINE, DEP DATE.

Inoltre è stata ottenuta un'altra variabile, T, come differenza tra DEP DATE e REQUEST DATE. Essa indica il tempo intercorrente (misurato in giorni) tra la richiesta di prezzo per il volo e il momento della partenza.
A questo punto, i dati sono stati riorganizzati ammettendo che a ogni modalità della variabile NEW corrispondesse una sola riga. In ognuna di esse, le variabili P1, P2, P3... riportano il prezzo osservato 1, 2, 3... giorni prima della partenza.
Si sarebbe potuto scegliere una discretizzazione più raffinata per le serie storiche dei prezzi dei voli (ad esempio, inserendo una rilevazione ogni 12 ore al posto di 24) ma ciò avrebbe portato un raddoppio delle osservazioni da trattare senza un significativo aumento di informazione, dato che raramente il prezzo di un volo cambia con frequenza maggiore di una volta al giorno, soprattutto a molti giorni di distanza dalla partenza. Ciò avrebbe anche reso le analisi più onerose dal punto di vista computazionale.
Naturalmente, però, nei dati in nostro possesso non vi è esattamente una rilevazione al giorno per ogni volo. Per questo motivo sono presenti molte celle vuote nella tabella. Per quanto riguarda i pochi casi in cui si avevano più osservazioni (non uguali) nell'arco della stessa giornata, è stata riportata la rilevazione di prezzo maggiore.

NEW	**P1**	**P2**	**P3**	**P4**	...
AGAORY AZ ...	45	39	39		...
BVSNAP I9 ...	99			79	...

Tabella 1.3: *Versione finale del database in cui ogni record contiene la serie storica osservata del prezzo di un volo. Oltre alle variabili mostrate sono presenti anche quelle che formano l'identificatore NEW.*

Prima di procedere con le analisi si sono effettuati alcuni altri accorgimenti.
Innanzi tutto, sono state escluse dal dataset tutte le serie storiche con meno di 30 osservazioni di modo da avere abbastanza informazione per ogni serie

storica. In secondo luogo, il tempo limite è stato fissato a 142 giorni prima della partenza (P143). Questa soglia è stata scelta con criterio soggettivo, dato che meno dell'1% dei voli presentava rilevazioni più in là nel tempo. Il database finale è composto da 1010 voli aerei e la percentuale di dati mancanti è del 72.38%.
In secondo luogo, per rendere i prezzi tutti confrontabili tra loro, si è scelto di dividere ognuno di essi per il prezzo minimo rilevato nella stessa riga.
Infine, si è trasformata la variabile DEP DATE in un indicatore del giorno della settimana in cui parte il volo.
Nelle tabelle 1.4 e 1.5 sono riportate alcune statistiche descrittive del dataset finale.

Min	Primo quartile	Mediana	Media	Terzo quartile	Max
35	36	38	39.49	42	59

Tabella 1.4: *Statistiche descrittive riguardanti il numero di rilevazioni presenti in ogni osservazione.*

DEP AIRPORT	DEP DATE	ARR AIRPORT	ID AIRLINE
BGY :270	1: 73	BVA :196	FR :795
BVA :172	2: 63	STN : 80	U2 : 85
CIA : 98	3:110	FEZ : 53	IV : 42
MRS : 67	4:160	GRO : 52	TO : 19
MXP : 65	5:308	OPO : 51	W6 : 19
BLQ : 58	6:212	RAK : 46	IG : 16
(Other):280	7: 84	(Other):532	(Other): 34

Tabella 1.5: *Statistiche descrittive riguardanti aeroporto e giorno della settimana di partenza, aeroporto di arrivo e compagnia aerea.*

Dalla tabella 1.5 si possono notare alcune caratteristiche che contraddistinguono il dataset finale. Le compagnie aeree più scelte sono quelle *low cost*, cosa abbastanza prevedibile dato che la convenienza è un aspetto molto importante per chi si rivolge a siti web per trovare voli aerei. Inoltre, la maggior parte dei voli di partenza è situata in Italia e Francia, il che ci fornisce informazioni di massima sulla provenienza degli utenti del sito web. Per quanto riguarda gli aeroporti di arrivo, escludendo la presenza degli stessi aeroporti presenti nella variabile DEP AIRPORT dovuta ai voli di ritorno, le destinazioni più scelte sono prevalentemente luoghi di vacanza nel Mediterraneo.

Capitolo 2

Il processo di Dirichlet

2.1 Introduzione

Come abbiamo già fatto notare, il progresso tecnologico degli ultimi decenni ha portato notevoli miglioramenti nelle possibilità computazionali a disposizione e questo si è tradotto, anche in statistica, in nuove metodologie a disposizione per l'analisi dei dati.
In particolare, la statistica Bayesiana ha visto allargare di molto i suoi orizzonti grazie a metodi di integrazione numerica basati su simulazioni Monte Carlo di catene di Markov (MCMC) che permettono di non limitarsi all'ambito delle famiglie coniugate per il calcolo delle distribuzioni *a posteriori*.
Una limitazione che invece non è superabile restando nell'ambito della statistica Bayesiana *parametrica* riguarda le possibili distribuzioni *a posteriori* raggiungibili, una volta impostata l'*a priori*. L'approccio parametrico, infatti, comporta il dover scegliere, non sempre con reali motivazioni, una specifica forma della distribuzione *a priori* e ciò limita lo spazio delle funzioni di densità *a posteriori* del parametro. Per fare un esempio, si pensi al caso del modello normale-normale per la stima del parametro di posizione. Qui tutte le funzioni di densità con supporto diverso da $(-\infty, +\infty)$, come anche le funzioni di densità asimmetriche o non unimodali, sono escluse dallo spazio di probabilità delle densità *a posteriori*.

Per superare questo limite, bisogna quindi poter assegnare probabilità non nulle a qualsiasi funzione *a priori* (o almeno a un insieme ragionevolmente più ampio di quello identificato in modo parametrico). Per far ciò, si può passare a considerare la statistica Bayesiana in un contesto non parametrico. Un tale approccio, applicato a problemi recenti, ha portato a risultati molto interessanti.
Nel prosieguo del capitolo verranno introdotti brevemente alcuni concetti di base della statistica Bayesiana non parametrica per passare velocemente alla presentazione dei modelli che verranno utilizzati per il trattamento dei dati di cui si è parlato nel capitolo precedente.

2.2 Il processo di Dirichlet e la modellazione di dati gerarchici

Un processo di Dirichlet è una distribuzione di probabilità definita sullo spazio delle funzioni di densità che induce una distribuzione di Dirichlet finita per ogni possibile partizione dei dati.
Senza scendere nei dettagli, questo processo, usato come distribuzione *a priori* nella stima bayesiana di una funzione di densità ha buone proprietà di flessibilità (possibilità di raggiungere qualsiasi distribuzione *a posteriori*) e coniugazione (la distribuzione *a posteriori* è ancora esprimibile come un processo di Dirichlet). Per la definizione di processo di Dirichlet e un'esposizione delle sue proprietà si veda Ferguson, 1974.
D'ora in poi verrà indicato con $DP(\alpha P_0)$ un processo di Dirichlet con P_0 misura di base, cioè la miglior scelta *a priori* per la distribuzione da stimare, e $\alpha > 0$ parametro di concentrazione, che può essere visto come la "fiducia" nella scelta di P_0.
Per capire meglio il significato di α e P_0 è utile introdurre la rappresentazione *stick breaking*, tramite cui se $P \sim DP(\alpha P_0)$, allora

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \ , \quad \theta_h \sim P_0 \ \ iid, \tag{2.1}$$

dove $\pi_h = V_h \prod_{l<h} (1 - V_l)$, con V_h osservazioni iid (indipendenti e identicamente distribuite) dalla distribuzione $Beta(1, \alpha)$ per $h = 1, 2, 3 \cdots + \infty$ e δ_θ è la misura di probabilità di Dirac su θ.
In pratica, il peso assegnato al primo atomo è una realizzazione della variabile aleatoria $Beta(1, \alpha)$, mentre la seconda realizzazione di questa variabile, V_2, denota la parte di probabilità non assegnata al primo atomo che deve essere allocata al secondo, e così via. La distribuzione congiunta della successione infinita $\{\pi_h\}$ così definita è detta distribuzione di Griffith, Engen e McCloskey (Pitman, 2002). In seguito si indicherà quanto appena descritto con $\{\pi_h\} \sim GEM(\alpha)$
Grazie a questa rappresentazione possiamo dare un significato più concreto a α. Se esso è molto piccolo, il valore atteso della distribuzione *Beta* è prossimo a uno e tutta la probabilità viene assegnata ai primi atomi. Se invece α tende a infinito, serve un numero elevato di atomi per avere una buona rappresentazione di P e si avrà $P \approx P_0$.
Un'applicazione immediata del processo di Dirichlet come distribuzione *a priori* è rappresentata dal seguente modello gerarchico:

$$y_{ij} = \mu_i + \epsilon_{ij}, \ \ \epsilon_{ij} \sim N(0, \sigma^2), \ \ \mu_i \sim P \tag{2.2}$$

dove le osservazioni sono raggruppate (con $i = 1, \ldots, k$ indice dei gruppi e $j = 1, \ldots, n_i$) e l'eterogeneità all'interno dei gruppi è data dalla componente di errore non prevedibile. Per quanto riguarda μ_i, è usuale la specificazione $\mu_i = \mu + \eta_i$, secondo cui la varietà tra i gruppi è stimata come discostamento dalla media complessiva μ. Si può porre, ad esempio, η_i come realizzazione di una variabile aleatoria normale di media nulla e questo comporta anche per μ_i una distribuzione normale, centrata sulla media campionaria $\hat{\mu}$.
Questa specificazione, però, è molto restrittiva. Essa comporta, infatti, code poco pesanti e quindi potrebbe non ammettere una variabilità adeguata tra i gruppi. Una possibile e immediata soluzione consiste nel considerare per

μ_i una distribuzione t di Student, ma anch'essa comporta delle restrizioni ingiustificate come quelle di unimodalità e simmetria.
Per non escludere nessuna possibilità dalla stima di P, si può adottare l'approccio non parametrico, ponendo come distribuzione *a priori* per P un processo di Dirichlet.
Oltre alle proprietà già citate riguardo questo approccio, è anche utile notare che la rappresentazione *stick-breaking* di un DP mostra la natura quasi certamente discreta di P, dalla quale si può ricavare una clusterizzazione (anche nel caso in cui essa non sia specificata dall'inizio nel modello). Per ogni possibile realizzazione di P (che verrà indicata con ψ), si ottiene un cluster contenente tutti i soggetti per cui $\mu_i = \psi$.
Questo metodo di clusterizzazione presenta un vantaggio importante rispetto a molte altre procedure di clustering. Essa infatti, non necessita di dover specificare *a priori* il numero di gruppi k in cui verrà suddiviso il campione.

2.2.1 Dati funzionali

La modellazione di dati funzionali può essere trattata con una diretta estensione del modello presentato al capitolo precedente.

$$y_i(t) = f_i(t) + \epsilon_i(t) \ , \quad \epsilon_i(t) \sim N(0, \sigma^2) \ \forall i, t \tag{2.3}$$

dove f_i è una funzione del tempo diversa per ogni osservazione, mentre $\epsilon_i(t)$ è la componente erratica per cui si considerano realizzazioni indipendenti per ogni osservazione e per ogni t. Se si pone $f_i \sim P$, P deve essere ora una distribuzione sullo spazio delle funzioni. La trattazione del problema in ambito statistico non parametrico è pertanto analoga a quella precedente, salvo il fatto che la misura base del processo di Dirichlet messo come *a priori* per P deve essere un processo stocastico. Una scelta comoda è il processo Gaussiano di cui vanno specificate media μ e funzione di autocovarianza C, $P_0 = GP(\mu, C)$.

In questo modo, nella rappresentazione *stick-breaking* gli atomi θ diventano realizzazioni di un processo di Gauss, e sono perciò funzioni. Come detto, la scelta $P_0 = GP(\mu, C)$ non è obbligatoria, ma comporta facilità computazionale dato che, per ogni fissato t, si ha che $f_i(t) = P(t)$ con $P(t) \sim DP(\alpha P_0(t))$ e con $P_0(t)$ distribuzione normale univariata. Considerando invece l'insieme (finito) di punti in cui viene stimata la funzione, il GP implica su di essi una distribuzione normale multivariata.

2.3 Applicazione ai dati

Il modello del paragrafo precedente può essere applicato ai dati in possesso (tralasciando momentaneamente la questione riguardante i dati mancanti). La sua specificazione completa è:

$$y_i(t) = f_i(t) + \epsilon_i(t) \ , \quad \epsilon_i(t) \sim N(0, \tau^{-1}) \ \ \forall i, t \tag{2.4}$$

dove è stato posto $\sigma^2 = \tau^{-1}$ per comodità di calcolo.
Resta da specificare la forma della funzione di covarianza C, per la quale si può scegliere una forma esponenziale del tipo:

$$C(t, t') = \kappa_1^{-1} e^{(-\kappa_2 |t - t'|)}. \tag{2.5}$$

E' ora possibile completare la specificazione del modello impostando le distribuzioni *a priori* per tutti i parametri del modello:

$\tau \sim Gamma(a_\tau, b_\tau)$
$\kappa_1 \sim Gamma(a_{\kappa_1}, b_{\kappa_1})$
$\kappa_2 \sim Gamma(a_{\kappa_2}, b_{\kappa_2})$
$f_i \sim P = \sum p_h \delta_{\theta_h}, \quad \theta_h \sim P_0 = GP(\mu, C)$
con $\{p_h\}_{h=1}^{\infty} \sim GEM(\alpha)$.

2.4 Una metafora utile: il ristorante cinese

Prima di passare alla verosimiglianza e allo studio della distribuzione *a posteriori* dei parametri, è utile introdurre altre proprietà del processo di Dirichlet che favoriscono l'interpretazione e la trattazione del modello. Il modello appena definito è scrivibile anche come un modello mistura:

$$y_i = \int N(y_i; \mu_i, \sigma^2) dP(f_i). \tag{2.6}$$

Per evitare di trattare l'elevato (e potenzialmente infinito) numero di parametri di P descritti nella rappresentazione *stick-breaking*, è possibile operare una marginalizzazione rispetto a P, considerando lo schema basato sulle urne di Polya introdotto da Blackwell e McQueen (1973) e ottenendo $(f_1, \ldots, f_n) \sim PU(\alpha P_0)$. Definendo con $f^{(i)}$ l'insieme di tutti i valori ottenuti per f senza considerare l'i-esima osservazione, è possibile ricavare la distribuzione *a priori* di ognuna delle funzioni di prezzo, date tutte le altre (distribuzione *full conditional*). Dopo la marginalizzazione, le osservazioni sono scambiabili, per cui la distribuzione è uguale per tutte:

$$p(f_i | f^{(i)}, \alpha) \propto \left(\frac{\alpha}{\alpha + n - 1}\right) P_0 + \left(\frac{1}{\alpha + n - 1}\right) \sum_{i' \neq i} \delta_{f_{i'}} \tag{2.7}$$

e dato che tutte le osservazioni che appartengono allo stesso gruppo hanno lo stesso effetto casuale, si può riscrivere la distribuzione come

$$p(f_i | \psi^{(i)}, \alpha) \propto \left(\frac{\alpha}{\alpha + n - 1}\right) P_0 + \left(\frac{1}{\alpha + n - 1}\right) \sum_{j=1}^{k^{(i)}} n_j^{(i)} \delta_{\psi_j^{(i)}}, \tag{2.8}$$

dove

- $\psi^{(i)}$ è l'insieme dei valori distinti di $f^{(i)}$
- $k^{(i)}$ è il numero di gruppi ottenuti senza considerare l'osservazione i
- $n_j^{(i)}$ è il numero di elementi per cui $f = f_j$, escludendo dal conto l'osservazione i

Torna qui utile introdurre la metafora del ristorante cinese dovuta a Pitman (1996), la quale descrive le probabilità marginali del processo di Dirichlet. Le unità statistiche sono viste come clienti che entrano, uno alla volta in un ristorante. Il primo cliente si siede al primo tavolo e sceglie il piatto del tavolo. In termini statistici, al primo gruppo (tavolo) viene assegnato il valore ψ_1 estratto da P_0.
Il secondo cliente ha due possibilità: sedersi al primo tavolo e mangiare il piatto già scelto dal primo cliente o sedersi al tavolo successivo e scegliere un nuovo piatto ψ_2.
In generale, l'allocazione dei clienti ai tavoli segue lo schema probabilistico a urne di Polya appena descritto per cui il cliente i sceglie un tavolo nuovo con probabilità proporzionale ad α e sceglie un tavolo già occupato proporzionalmente al numero di clienti che occupano già il tavolo (n_j).
Perciò, se α è molto grande, si ottengono tanti gruppi quanti sono le osservazioni nel campione (idealmente si ha $f_i \sim P$, per cui si ricade nel modello gerarchico parametrico). D'altra parte, se si pone $\alpha = 0$, si ottiene un unico gruppo e quindi viene stimato un modello senza eterogeneità ($f_i = f \ \forall \, i$). Più precisamente, è possibile ottenere la distribuzione del numero di gruppi k in funzione di α:

$$p(k = m|\alpha, n) = c_n(m) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \tag{2.9}$$

con $c_n(m) = p(k = m|\alpha = 1, n)$ (Antoniak,1974). Da essa, si può ricavare il valore atteso *a priori* del numero di gruppi k che è proporzionale a $\alpha log(n)$. Si otterrà quindi un numero di gruppi che cresce molto lentamente con l'aumentare del numero di osservazioni.

2.5 Verosimiglianza e distribuzione *a posteriori*

Per quanto riguarda i dati, si hanno n insiemi di osservazioni, ognuno dei quali è costituito da T misurazioni della funzione in spazi temporali pre-

definiti.
Definita ogni osservazione $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, la verosimiglianza associata ai dati è

$$L(y|f,\tau) \propto \tau^{\frac{nT}{2}} exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - f_i)'(y_i - f_i)\right) \tag{2.10}$$

E' poi possibile esplicitare il modello, utilizzando l'indicatore S_i, dove $S_i = h$ vuol dire che l'osservazione i-esima fa parte del cluster h.

$$(y_i|S_i = h) \sim N_T(\theta_h, \tau^{-1}) \tag{2.11}$$

Prima di ricavare la distribuzione *a posteriori*, è bene definire chiaramente la simbologia che verrà utilizzata.
Ognuna delle n funzioni (dati) viene stimata con una delle k che caratterizzano i gruppi. Si indicherà con ϕ_i ognuna delle funzioni stimate e con ψ_j i k distinti valori di ϕ, per cui $S_i = h$ implica che $\phi_i = \psi_h$. Le lettere senza indici, indicheranno insiemi di osservazioni: $\phi = \{\phi_i,\ i = 1, \ldots, n\}$ e $\psi = \{\psi_j,\ j = 1, \ldots, k\}$. Il vettore $S = (S_1, \ldots, S_n)$ indicherà l'allocazione di ognuna delle osservazioni.
Infine la dicitura $...^{(i)}$, indicherà la grandezza o il vettore, calcolati senza considerare l'i-esima osservazione, ad esempio $\psi^{(i)}$ è l'insieme di tutti i possibili valori di ϕ presi una sola volta, e può essere uguale a ψ se $S_i \neq k^{(i)} + 1$.
Ricordando che $(\phi|\alpha, \kappa_1, \kappa_2) \sim PU(\alpha, GP(\mu, C))$, possiamo ottenere la distribuzione *full conditional* di ϕ_i che è uguale per ogni i dato che, dopo l'integrazione rispetto alla misura base, le osservazioni sono interscambiabili. Essa dipenderà da $S^{(i)}, k^{(i)}$ e $\psi^{(i)}$, oltre che da α.

$$(\phi_i|y_i, S^i, k^i, \psi^i, \alpha) \sim q_{i0}P_{i0} + \sum_{j=1}^{k^{(i)}} q_{ij}\delta_{\psi_j^{(i)}} \tag{2.12}$$

con

$$P_{i0} = \frac{L_i(y_i|\phi_i)P(\phi_i)}{\int L_i(y_i|\phi_i)P(\phi_i)d\phi_i} \tag{2.13}$$

I pesi della mistura sono invece espressi da:

$$q_{ij} \propto \begin{cases} \alpha h_i(y_i) \\ n_j^{(i)} L(y_i|\psi_j) \end{cases} \tag{2.14}$$

dove

$$\begin{aligned} h_i(y_i) &= \int L_i(y_i|\psi_j) dP_0(f_i) \\ &= \int N_T(y_i; f_j, \tau^{-1} I_T) N_T(f; \mu, C) df \\ &= N_T(y_i, \mu, \tau^{-1} I_T + C) \end{aligned}$$

Le funzioni di prezzo f_i possono essere estratte per ciascun soggetto, ma ciò richiederebbe molto tempo per ogni iterazione. E' invece più conveniente estrarre prima l'indicatore di gruppo S, e poi ψ. La distribuzione *full conditional a posteriori* di S_i è data da:

$$p(S_i = j|y_i, S^{(i)}, k^{(i)}, \psi^{(i)}) = q_{ij} \quad i = 1, \ldots, n. \tag{2.15}$$

Se $S_i = 0$, l'osservazione va a inaugurare un nuovo cluster $k = k^{(i)} + 1$.
Dopo questo passo, si può aggiornare ψ dalla sua distribuzione *full conditional*. Detta p_0 la densità della distribuzione $N_T(\mu, C)$, si ha che

$$\begin{aligned} p(\psi_j|S, y) &\propto p_0(\psi_j) \prod_{i:S_i=j} L_i(y_i) \\ &\propto N_T([\tau I + m_j C^{-1}]^{-1}[\tau I \mu + m_j C \bar{y}], [\tau I + m_j C^{-1}]^{-1}) \end{aligned}$$

con m_j numero di osservazioni presenti nel cluster j e $\bar{y}$ media di queste osservazioni.

2.6 Algoritmo

Per il calcolo della distribuzione *a posteriori* bisogna utilizzare algoritmi di tipo MCMC. In particolare, esistono due vie.

L'algoritmo *blocked Gibbs sampler* (Ishwaran e James, 2001) prevede di usare solo i primi N termini della rappresentazione *stick-breaking*, ponendo $V_N = 1$. In questo modo N diventa una soglia massima per il numero di cluster.
La versione (MacEachern, 1994), invece, usa lo schema delle urne di Polya presentato precedentemente anche tramite la metafora del ristorante cinese.
Quando non si hanno informazioni *a priori* su N, è conveniente dal punto di vista computazionale utilizzare il secondo algoritmo piuttosto che porre una soglia molto grande per il numero di gruppi. Per questo motivo, verrà ora presentato per punti lo schema dell'algoritmo *collapsed Gibbs sampler*.

1. L'inizializzazione consiste nell'estrazione casuale dei parametri dalla loro distribuzione *a priori*. Si alloca poi di default la prima osservazione al primo cluster.

2. Per ogni osservazione bisogna calcolare i pesi q_{ij} ed estrarre poi l'indicatore di gruppo S_i. Questo passo deve essere ripetuto per ogni osservazione.

3. E' ora possibile aggiornare la traiettoria ψ_j per ogni gruppo dalla distribuzione *full conditional*.

4. Infine, si devono aggiornare i parametri τ, κ_1 e κ_2 dalle loro distribuzioni *a posteriori*:

$$p(\tau|y, \psi, S) = Gamma\left(a_\tau + \frac{nT}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{n}(y_i - \phi_i)'(y_i - \phi_i)\right). \tag{2.16}$$

   Per quanto riguarda i parametri della funzione di autocovarianza, essi non hanno distribuzione in forma chiusa ed è pertanto necessario annidare nell'algoritmo un passo Metropolis-Hastings:

$$p(\kappa_1, \kappa_2| \ldots) \propto |C|^{-\frac{1}{2}} e^{\left(-\frac{1}{2}(\phi_i-\mu)'C^{-1}(\phi_i-\mu)\right)} \kappa_1^{a_{\kappa_1}-1} \kappa_2^{a_{\kappa_2}-1} e^{-b_{\kappa_1}\kappa_1 - b_{\kappa_2}\kappa_2}. \tag{2.17}$$

Un aspetto di cui non si è ancora parlato è dato dalla trattazione dei dati mancanti. Fortunatamente, non ci sono grosse complicazioni. Più precisamente, le uniche implicazioni si hanno nei passi 2 e 3 dell'algoritmo.
Nel calcolo dei pesi q_{ij} la verosimiglianza L_{y_i} deve essere calcolata solamente considerando i tempi per cui si hanno osservazioni e ugualmente si deve procedere per l'estrazione dalla distribuzione $h_i(y_i)$.
Una volta ottenuti i gruppi, poi, l'aggiornamento della funzione di prezzo che li caratterizza deve essere fatta solo nei tempi per cui ci sono osservazioni in quel gruppo, e solo in base alle osservazioni che contengono dati validi per ogni t.
Ad esempio, se in un cluster ci sono solo osservazioni con dati validi solo per $t \in (1, 20)$, sarà impossibile che a questo gruppo siano aggiunte funzioni per cui si sono osservati valori per t maggiori. La traiettoria ψ sarà poi ottenuta solo per i t corrispondenti, creando una clusterizzazione basata non esclusivamente sui valori osservati, ma anche sui tempi in cui sono stati richiesti i preventivi per i voli.

Capitolo 3

Oltre il processo di Dirichlet

Con i risultati del modello esposto al capitolo 2, si è potuta operare una clusterizzazione dei voli e studiare così alcune caratteristiche generali comuni alle osservazioni dello stesso gruppo. Le considerazioni fatte possono essere un buon punto di partenza, ma non sono sufficienti se si vuole una stima accurata di uno dei valori rilevati in una specifica funzione di prezzo. Una delle ragioni è da ricercarsi nel fatto che, a ogni iterazione dell'algoritmo, i valori predetti della funzione sono ricavati dalla distribuzione che caratterizza il cluster e sono pertanto uguali per tutte le osservazioni del gruppo. Una previsione fatta in questo modo è spesso poco precisa.
Per questo motivo, in questo capitolo verrà introdotto il processo di Dirichlet gerarchico il quale può essere utilizzato applicandolo a una funzione di prezzo alla volta, per ottenere una previsione accurata del suo andamento nel tempo. Per far ciò si utilizzerà la struttura probabilistica delle catene di Markov nascoste (Rabiner, 1989). Il modello che ne risulta viene detto processo di Dirichlet gerarchico con catene di Markov nascoste (Teh, Jordan, Beal e Blei, 2006) che verrà abbreviato con l'acronimo HDP-HMM.
Inoltre, data la struttura dei dati, con prezzi che spesso si ripropongono uguali per più giorni, si utilizzerà la versione *sticky* HDP-HMM introdotta da Fox (2010) che permette di superare alcuni limiti del modello.
Nel primo paragrafo si introdurrà il modello di Dirichlet gerarchico, nel se-

condo esso verrà applicato alla struttura probabilistica delle catene di Markov nascoste e nel terzo verrà presentato il modello *sticky* HDP-HMM. Nell'ultima parte del capitolo si parlerà invece dell'algoritmo di calcolo utilizzato per implementare il modello.

3.1 Il modello di Dirichlet gerarchico

Non sarà sfuggito, che la stima di P_0 del modello precedente, è stata trattata in modo parametrico, predeterminandone cioè in qualche modo la forma tramite μ, κ_1, κ_2.
Un'altra possibilità è data invece dall'estrarre anche P_0 da una distribuzione il cui supporto è un insieme di misure di probabilità.
Uno dei modi per trattare non parametricamente anche questa parte del modello, consiste nel processo di Dirichlet gerarchico (HDP).
L'idea principale consiste nell'estrarre la misura base di un processo di Dirichlet da un altro processo di Dirichlet. Più in generale, è possibile collegare tra loro un insieme di DP nel seguente modo:

$$P_0 \sim DP(\gamma, Q)$$
$$P_j \sim DP(\alpha, P_0)$$

dove j è l'indicatore di gruppo e Q è una distribuzione in $\Re$. In questo modo, tutti i gruppi condividono tra loro lo stesso insieme di atomi determinato da P_0.
Per favorire l'interpretazione del processo e della natura del collegamento tra i vari DP, è utile ora introdurre la sua rappresentazione *stick-breaking*, così come fatto in precedenza.
Per prima cosa, è subito possibile scrivere:

$$P_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{3.1}$$

con $\{\beta_k\} \sim GEM(\gamma)$ e θ_k estrazioni iid da Q.
Ora, dato che anche tutti i P_j sono dei DP, è possibile ottenere per ognuno di essi la rappresentazione *stick-breaking*. Dato che il supporto di P_j è necessariamente contenuto in quello di P_0, essa è una somma pesata degli stessi atomi che compongono P_0, con pesi diversi da gruppo a gruppo:

$$P_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}. \tag{3.2}$$

Ciò che resta da trovare è la relazione tra $\beta = \{\beta_k\}_{k=1}^{\infty}$ e $\pi_j = \{\pi_{jk}\}_{k=1}^{\infty}$. Usando la definizione di DP di Ferguson (1973), si può ottenere

$$\pi_j \sim DP(\alpha, \beta) \tag{3.3}$$

dove qui l'insieme dei pesi β individua una distribuzione di probabilità discreta sugli atomi θ_k. Dalla 3.3 è possibile ottenere in modo esplicito la costruzione di π_j condizionata a β:

$$\pi_{jk} = V_{jk} \prod_{l=1}^{k-1} (1 - V_{jl}) \tag{3.4}$$

per ogni j e per $k = 1, ..., \infty$, dove

$$V_{jk} | \alpha, \beta \sim Beta\left(\alpha\beta_k, \alpha\left(1 - \sum_{l=1}^{k} \beta_l\right)\right). \tag{3.5}$$

3.1.1 Un franchising di ristoranti cinesi

Così come il processo del ristorante cinese descrive le probabilità marginali del processo di Dirichlet, esiste una rappresentazione analoga valida per l'HDP chiamata *Chinese restaurant franchise* (CRF, Teh et al., 2006).
In questa nuova metafora, invece di un solo ristorante, si ha a disposizione un intero set di locali indicizzati da j. In ognuno di questi locali, i clienti scelgono i tavoli in modo completamente analogo a quello del CRP, e ciò avviene in modo indipendente da ciò che accade negli altri ristoranti.
La relazione presente tra i ristoranti è data dalla condivisione dello stesso

menu (gli stessi atomi che definiscono P_0).
E' possibile ottenere la rappresentazione tramite urne di Polya anche per questo processo, ma per farlo è necessario ridefinire la simbologia utilizzata.
Sia ora θ_{ji} una variabile aleatoria che descrive la probabilità che l'i-esimo cliente scelga il ristorante j, distribuita secondo P_j. Sia poi θ^*_{jt} un'altra variabile aleatoria che indica la scelta del tavolo t nel ristorante j distribuita stavolta secondo P_0. Infine, i piatti scelti si distribuiscono secondo la misura di base Q e sono indicati con θ^{**}_k. Ogni cliente sceglie un tavolo e ogni tavolo serve un solo piatto.
Con t_{ji} si indicherà il tavolo scelto dal cliente i nel ristorante j mentre con k_{jt} il piatto servito dal tavolo t nel ristorante j. Si ha che $\theta_{ji} = \theta^*_{jt_{ji}} = \theta^{**}_{k_{jt_{ji}}}$.
Da ultimo, sia n_{jtk} il numero di clienti che mangiano il piatto k, seduti al tavolo t del ristorante j e m_{jk} il numero di tavoli che servono il piatto k nel ristorante j. K indicherà il numero totale di piatti serviti in tutto il franchise mentre i puntini nei pedici indicheranno invece i conteggi marginali. Ad esempio, $n_{j\cdot k}$ è il numero di clienti che mangiano il piatto k nel j-esimo ristorante indipendentemente dal tavolo al quale sono seduti.
Integrando rispetto a P_j, si può ottenere la distribuzione condizionata di θ_{ji} in forma di schema a urne di Polya:

$$\theta_{ji} | \left\{\theta_{jh}\right\}_{h=1}^{i-1}, \alpha, P_0 \sim \frac{\alpha}{\alpha + n_{j\cdot\cdot}} P_0 + \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{\alpha + n_{j\cdot\cdot}} \delta_{\theta^*_{jt}}. \tag{3.6}$$

Anche in questo caso, il cliente si siede a un nuovo tavolo con probabilità proporzionale a α e sceglie un tavolo già occupato in base al numero di persone che lo hanno scelto in precedenza.
Dato che anche θ^*_{jt} è generata da osservazioni iid da P_0, possiamo ottenere una distribuzione simile a quella precedente, integrando rispetto a questa misura base:

$$\theta^*_{jt} | \theta^*_{prec}, \gamma, Q \sim \frac{\gamma}{\gamma + m_{\cdot\cdot}} Q + \sum_{k=1}^{K} \frac{m_{\cdot k}}{\gamma + m_{\cdot\cdot}} \delta_{\theta^{**}_k}, \tag{3.7}$$

dove con θ^*_{prec} si intendono tutti i θ^* che hanno avuto realizzazione precedente a quella di θ^*_{jt} (in altre parole si ha che l'indice t va da 1 a $m_{j\cdot}$ per ogni j).
In questo caso, si nota come la scelta del piatto non dipenda dal numero di volte che il piatto è stato scelto in un ristorante, ma dal numero totale di consumatori che stanno mangiando quel piatto in tutto il franchise ($m_{\cdot k}$).

3.2 Catene di Markov nascoste

Una catena di Markov nascosta è un processo secondo cui una serie storica $\{x_t\}$ è generata da una sottostante catena di Markov $\{\theta_t\}$ e da una funzione di emissione probabilistica f tale che $f(\theta_t) = x_t$.

$$\begin{aligned} \theta_t|\theta_{t-1}, P_{\theta_{t-1}} &\sim P_{\theta_{t-1}} \\ x_t|\theta_t &\sim F_{\theta_t} \qquad con \quad \theta_t \in \Theta \end{aligned}$$

e Θ insieme di possibili stati predefiniti.
Si può applicare il processo di Dirichlet gerarchico al modello di Markov nascosto (HMM) di modo da non dover specificare inizialmente quanti sono gli stati che si possono presentare nella catena sottostante e di ammetterne così un numero potenzialmente infinito.
Si consideri infatti la struttura del processo generatore dei dati supposto. Noto θ_t, l'osservazione x_{t+1} viene generata scegliendo prima lo stato successivo della catena di Markov θ_{t+1} secondo P_{θ_t}. Dato che l'insieme dei possibili stati validi per θ_{t+1} deve essere potenzialmente uguale per ogni possibile modalità di θ_t, mentre la probabilità di transizione devono essere specifiche per ognuna di queste modalità, è ragionevole pensare di impostare un HDP in cui le probabilità di transizione $P_{\theta_j} = P(\theta_{t+1}|\theta_t = \theta_j)$ sono modellate secondo processi di Dirichlet che hanno tutti in comune la stessa misura base:

$$
\begin{aligned}
P_0 &\sim DP(\gamma, Q) \\
P_{\theta_j} &\sim DP(\alpha, P_0) \qquad con \quad \theta_j \in \Theta,
\end{aligned}
$$

dove, in questo caso, Θ è determinato da P_0. Il modello probabilistico appena descritto prende il nome di processo di Dirichlet gerarchico con catena di Markov nascosta (HDP-HMM). Anche per esso è utile vedere la rappresentazione *stick-breaking*, mantenendo la notazione θ^{**} per gli atomi della misura di base Q:

$$
\begin{aligned}
P_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}, \\
P_{\theta_l^{**}} &= \sum_{k=1}^{\infty} \pi_{\theta_{lk}^{**}} \delta_{\theta_k^{**}} \qquad per \quad l = 1, \ldots, \infty,
\end{aligned}
$$

con

$$
\begin{aligned}
\theta_k^{**} | Q &\sim Q, \\
\beta | \gamma &\sim GEM(\gamma), \\
\pi_{\theta_k^{**}} | \alpha, \beta &\sim DP(\alpha, \beta) \qquad k = 1, \ldots, \infty
\end{aligned}
$$

Con notazione un po' più snella si può indicare con l'intero k lo stato θ_k^{**} e con z_t la variabile che indica lo stato al tempo t. In questo modo si ottiene che $\theta_t = \theta_k^{**} \leftrightarrow z_t = k$ e si può scrivere $\pi_{\theta_k^{**}} = \pi_k$. La struttura markoviana si può ora esprimere con:

$$
\begin{aligned}
z_t | z_{t-1}, \pi_{z_{t-1}} &\sim \pi_{z_{t-1}} \\
x_t | z_t, \theta_{z_t}^{**} &\sim F_{\theta_{z_t}^{**}}
\end{aligned}
$$

In questo modo si può notare chiaramente che l'HDP-HMM può essere interpretato come una catena di Markov nascosta in cui il numero di stati è potenzialmente infinito. Le differenti probabilità di transizione sono governate ognuna da un suo DP, e collegate tra loro tramite la struttura gerarchica del CRF.
Come accennato nell'introduzione, questo modello verrà applicato a una singola funzione di prezzo. A ogni possibile stato presente nella catena, corrisponde una parametrizzazione della funzione di emissione e una probabilità di transizione allo stato successivo. In altre parole, l'insieme degli stati possibili corrisponde all'insieme dei ristoranti presenti nel CRF.
Nel caso dei dati relativi ai prezzi dei voli aerei, ogni stato corrisponde un determinato prezzo base e la funzione di emissione permette di modellare alcune variazioni del prezzo rilevato non dovute a scelte della compagnia aerea (come ad esempio l'utilizzo di promozioni).
Per quanto riguarda la trattazione dei dati mancanti, in fase di applicazione si supporrà che in corrispondenza della mancanza di rilevazioni lo stato della catena sia uguale a quello precedente (che nel tempo reale è invece quello successivo, per come sono stati costruiti i dati).

3.3 Problema e soluzione

Come si è visto nella costruzione della rappresentazione *stick-breaking*, nell'HDP si ha che $\pi_j|\alpha,\beta \sim DP(\alpha,\beta)$. Nella versione HDP-HMM, a ogni π_j corrisponde uno degli stati della catena di Markov sottostante e quindi una diversa probabilità di transizione allo stato successivo.
Questa struttura, porta gli stati ad avere simili distribuzioni di transizione dato che $E\left[\pi_{jk}|\left\{\beta_k\right\}_{k=1}^{\infty}\right] = \beta_k$ ma non riesce a distinguere periodi di persistenza ($z_t = z_{t-1}$) da passaggi da uno stato all'altro.
Quando si cerca di modellare dati caratterizzati da elevata persistenza degli stati nel tempo, la natura flessibile dell'HDP-HMM porta alla stima di catene di Markov che oscillano velocemente tra stati ridondanti. Questo problema

può essere accentuato se si pensa alla possibilità di avere una funzione di emissione multimodale. In questo caso, è più facile scindere le osservazioni in gruppi in modo da avere per ogni moda uno stato diverso, permettendo rapidi spostamenti da uno stato all'altro. Tali irrealistiche situazioni hanno alta probabilità *a posteriori* dato che non c'è un'adeguata penalità per il numero di stati stimati.
Tutto ciò può non costituire un problema se gli stati sono considerati esclusivamente come variabili di supporto e si è interessati solo alla sequenza $\{x_t\}$, ma non è affatto desiderabile se si vuole fare inferenza sul numero di stati o sulle probabilità di auto-transizione.

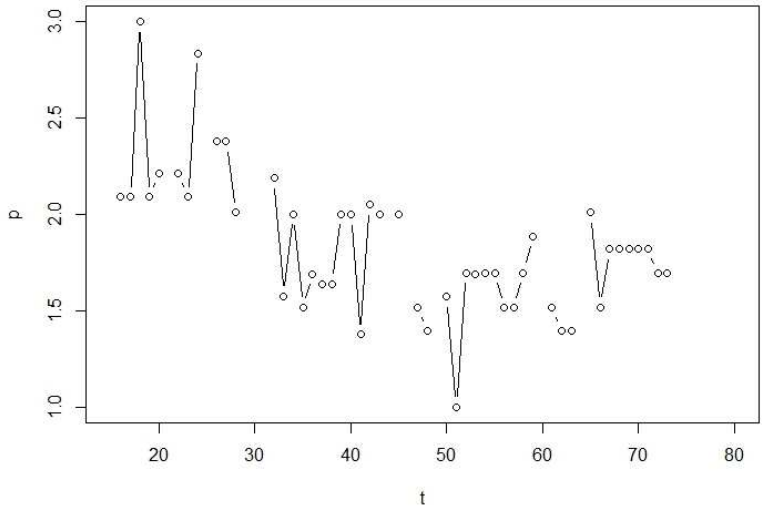

Figura 3.1: *Funzione di prezzo osservata del volo Bergamo-Sharm El Sheik del 5/9/2012. Si notino i periodi di persistenza, in particolare all'inizio (destra) della catena.*

Nel caso delle funzioni di prezzo dei voli aerei, la persistenza è relativamente elevata ed è inoltre interesse dell'azienda cercare di prevedere il momento in cui si assisterà a un cambiamento dello stato della catena dato che esso cor-

risponde a un cambiamento del prezzo atteso per il giorno successivo. Per questi motivi è necessario superare questo limite che caratterizza il modello HDP-HMM.
Per risolvere questo problema Fox, et al. (2008) hanno proposto una variazione nella struttura dei π_j, introducendo un nuovo parametro $\xi > 0$ che aumenta la probabilità di transizione allo stato attuale. Mantendendo fisso $\beta|\gamma \sim GEM(\gamma)$, si pone:

$$\pi_j|\alpha, \beta, \xi \sim DP\left(\alpha + \xi, \frac{\alpha\beta + \xi\delta_j}{\alpha + \xi}\right). \tag{3.8}$$

Se $\xi = 0$ si torna al modello precedente. Questa generalizzazione del modello viene definita *sticky* e si parla quindi di *sticky* HDP-HMM.
Anche questa estensione trova il suo corrispettivo nella metafora del ristorante cinese. Nel franchising di ristoranti cinesi valido per l'HDP-HMM, viene introdotta la fedeltà dei clienti.
In questo nuovo CRF, ogni ristorante ha una specialità della casa che viene indicizzata con lo stesso indice valido per i ristoranti. La specialità della casa di ogni ristorante è servita anche in ognuno degli altri ristoranti della catena, ma si può dire che la sua versione è migliore rispetto a quella degli altri ristoranti, pertanto la probabilità di sceglierla è maggiore rispetto al solito. Infatti, si può notare che il piatto di ogni tavolo è estratto da:

$$k_{jt}|\alpha, \beta, \xi \sim \frac{\alpha\beta + \xi\delta_j}{\alpha + \xi}. \tag{3.9}$$

Più semplicemente, ogni ristorante è caratterizzato da un diverso insieme di probabilità relative ai piatti del menu, con un aumento della probabilità di scegliere la specialità della casa.
Si ricordi ora che ogni cliente sceglie prima il ristorante, poi il tavolo ed eventualmente il piatto (se è il primo cliente del tavolo). Per la struttura markoviana, si ha che z_t determina lo stato z_{t+1}. In altre parole, vedendo il cliente t-esimo come genitore e il t+1-esimo come figlio, si può dire che la scelta del primo influenza quella del secondo. Più precisamente, il genitore sceglie un

ristorante (dipendentemente da z_{t-1}), sia esso j. Una volta entrato nel ristorante, egli sceglie il tavolo e di conseguenza il piatto, e come visto, l'evento più probabile è che il genitore consumi la specialità della casa. Date le nuove probabilità di transizione, per il figlio sarà più probabile la scelta dello stesso ristorante (e quindi dello stesso piatto) del padre. Si genera così una specie di fedeltà di famiglia.
Si supponga ora, che il genitore scelga un piatto che non è la specialità della casa. L'evento più probabile per il figlio è ora dato dalla scelta del ristorante che ha come specialità della casa il piatto scelto dal padre. Si può quindi dire che la fedeltà generazionale è data dal fatto che i figli hanno gli stessi gusti del proprio genitore. Questo meccanismo si traduce in un'elevata persistenza degli stati, così come desiderato.
Per semplificare l'algoritmo, sarà utile introdurre un'altra variabile. Ciò viene fatto qui per completare la metafora.
Nella scelta del piatto è più comodo impostare il seguente meccanismo:

$$\begin{aligned}
&\bar{k}_{jt}|\beta \sim \beta \\
&\omega_{jt}|\alpha,\xi \sim Ber(\rho), \qquad con \quad \rho = \frac{\xi}{\alpha+\xi} \\
&k_{jt}|\bar{k}_{jt},\omega_{jt} = \begin{cases} \bar{k}_{jt} & se \quad \omega_{jt}=0 \\ j & se \quad \omega_{jt}=1 \end{cases}
\end{aligned}$$

Si può considerare $\bar{k}$ il piatto che il cliente sceglie, senza conoscere quale sia la specialità della casa. Per questo motivo, la scelta di $\bar{k}$ è assolutamente uguale a quella di k nel processo *non-sticky*. La variabile ω è invece considerabile come variabile di rivalutazione della scelta. Si può supporre che al momento della scelta il cameriere consigli al cliente di scegliere la specialità della casa e che questo consiglio possa essere accettato o meno, con ξ che diventa quindi un indicatore della 'capacità di convincimento del cameriere'. Analogamente al modello *non-sticky* k resta il piatto effettivamente servito al cliente.
La distribuzione in forma di schema a urne di Polya è la seguente:

$$p\left(t_{ji}|\left\{t_{jl}\right\}_{l=1}^{i-1},\alpha,\xi\right) \propto (\alpha+\xi)P_0 + \sum_{t=1}^{m_{j\cdot}} n_{jt\cdot}\delta_{t_{jt}}$$

$$p\left(\bar{k}_{jt}|\bar{k}_{prec},\gamma\right) \propto \gamma Q + \sum_{k=1}^{\bar{K}} \bar{m}_{\cdot k}\delta_{\bar{k}_{jt}}$$

dove la simbologia è del tutto analoga a quella usata per l'HDP, con $\bar{K}, \bar{m}\ldots$ che si riferiscono alle rispettive grandezze considerate prima che intervenga la variabile ω.

3.4 Algoritmo

Fox, Sudderth, Jordan e Willsky (2008) hanno dimostrato che la versione *sticky* dell'HDP-HMM, non solo riesce a catturare bene processi caratterizzati da elevata persistenza di stati, ma non incorre nemmeno in difficoltà predittive quando invece il loro andamento è molto variabile. Esso è inoltre ben adattabile al caso in cui le funzioni di emissioni siano di natura multimodale.
Nel loro lavoro, sono state presentate due versioni per l'algoritmo di calcolo: il campionamento per assegnazione diretta e il campionamento in blocco delle sequenze degli stati. Per entrambe le versioni dell'HDP-HMM, alcune simulazioni hanno dimostrato che il secondo riesce a dare una stima più precisa della vera sequenza degli stati, pertanto di seguito verrà proposto solamente questo. Ci si limiterà inoltre alla versione dell'algoritmo con funzioni di emissione gaussiane, dato che nel caso preso in esame, non ha senso considerare strutture più complesse. Per tutti gli approfondimenti, si rimanda al lavoro citato a inizio paragrafo.
L'algoritmo è un adattamento della procedura forward-backward per le HMM introdotta da Rabiner (1989). Per poterlo implementare bisogna usare un'approssimazione delle probabilità di transizione (dato che sono caratterizzate

da un numero potenzialmente infinito di stati). Tra i metodi proposti, quello di più semplice realizzazione impone di considerare l'approssimazione proposta da Ishwaran e Zarepour (2002) per il processo di Dirichlet, data da

$$GEM_L(\alpha) = Dir(\alpha/L, \ldots, \alpha/L),$$

in cui L è numero intero ragionevolmente maggiore del vero numero di stati della catena di Markov.
Per motivi computazionali, inoltre, è bene considerare la seguente relazione deterministica

$$\begin{aligned} \alpha &= (1-\rho)(\alpha+\xi) \\ \xi &= \rho(\alpha+\xi) \end{aligned}$$

che permette di riparametrizzare, considerando ρ e $(\alpha + \xi)$ come parametri al posto di α e ξ.

3.4.1 Campionamento di β e π

L'approssimazione per il processo di Dirichlet appena presentata consente di elicitare la distribuzione *a priori* di β:

$$\beta|\gamma \sim Dir(\gamma/L, \ldots, \gamma/L).$$

Mentre per la distribuzione *a priori* delle probabilità di transizione si ha

$$\pi_j|\beta, \alpha, \xi \sim Dir(\alpha\beta_1, \alpha\beta_2, \ldots, \alpha\beta_j + \xi, \ldots, \alpha\beta_L).$$

Le rispettive distribuzioni *a posteriori* sono

$$\begin{aligned} \beta|\gamma, \bar{m} &\sim Dir(\gamma/L + \bar{m}_{\cdot 1}, \ldots, \gamma/L + \bar{m}_{\cdot L}) \\ \pi_j|\beta, \alpha, \xi, \{z\} &\sim Dir(\alpha\beta_1 + n_{j1}, \ldots, \alpha\beta_j + \xi + n_{jj}, \ldots, \alpha\beta_L + n_{jL}) \end{aligned}$$

Conviene qui richiamare che con n_{jk} si intende il numero di transizioni dallo stato (o ristorante) j allo stato k e con $\bar{m}_{jk}$ il numero di tavoli del ristorante j che hanno scelto il piatto k, prima dell'intervento del cameriere ω.

3.4.2 Variabili ausiliarie e sequenza degli stati

Per le variabili ausiliarie si ha la seguente distribuzione congiunta:

$$p(m, \omega, \bar{m} | \{z\}, \beta, \alpha, \xi) = p(\bar{m}|m, \omega, \{z\}, \beta, \alpha, \xi) \times \\ \times p(\omega|m, \{z\}, \beta, \alpha, \xi) p(m| \{z\}, \beta, \alpha, \xi).$$

Si consideri la distribuzione di m. Si può pensare che, poiché si conosce la sequenza degli stati $\{z\}$, si conosce anche la suddivisione dei clienti per ristoranti e per piatti consumati e quindi l'unico fattore di incertezza è dato dal fatto che non si sa quanti tavoli stanno servendo lo stesso piatto. Perciò, per ottenere un campionamento per m, bisogna ottenere la distribuzione dei tavoli, nota quella dei piatti. Si può mostrare che essa segue un DP con parametro di concentrazione pari a $\alpha\beta_k + \xi\delta(k, j)$:

$$p(t_{ij} = t | k_{ij} = k, t^{(ij)}, k^{(jt)}, \{x\}, \beta, \alpha, \xi) \propto \begin{cases} n_{jt}^{(ji)} & t \in \{1, \ldots, T_j\} \\ \alpha\beta_k + \xi\delta(k, j) & t = T_j + 1 \end{cases}$$

Per quanto riguarda ω si hanno due casi. Se $k \neq j$, allora si hanno esattamente m_{jk} tavoli con $\omega_{jt} = 0$. Per l'altro caso invece, si può supporre di conoscere il piatto considerato $\bar{k}$ e poi integrare su tutti i possibili suoi valori.

$$p(\omega_{jt} | k_{jt} = j, \beta, \rho) \propto \begin{cases} \beta_j(1 - \rho) & \omega_{jt} = 1 \\ \rho & \omega_{jt} = 0. \end{cases}$$

Infine, $\bar{m}_{jk}$ si può ottenere in modo deterministico dato che $\bar{m}_{jk} = m_{jk}$ se $j \neq k$ e $\bar{m}_{jj} = m_{jj} - \omega_{j\cdot}$, con $\omega_{j\cdot}$ che segue la distribuzione binomiale.
Per quanto riguarda la sequenza degli stati, essa può essere ottenuta stato per stato, data la struttura markoviana della catena.

$$p(\{z\} \mid \{x\}, \pi, \theta^{**}) = \prod_{h=1}^{T} p(z_h | z_{h-1}, \{x\}, \pi, \theta^{**})$$

Partendo da z_1, si ha che

$$p(z_1 | \{x\}, \pi, \theta^{**}) \propto p(z_1) f(x_1 | \theta^{**}_{z_1}) m_{2,1}(z_1),$$

dove $m_{t,t-1}(z_{t-1}) \propto p(\{x\} | z_{t-1}, \pi, \theta)$ è detto 'messaggio all'indietro' da z_t a z_{t-1} ed è interpretabile come l'informazione proveniente dallo stato successivo (per approfondimento si veda Rabiner, 1989).
La distribuzione del generico elemento della catena è invece data da

$$p(z_t | z_{t-1} \{x\}, \pi, \theta^{**}) \propto p(z_t | \pi_{z_{t-1}}) f(x_t | \theta^{**}_{z_t}) m_{t+1,t}(z_t).$$

Nel prossimo paragrafo, viene ricapitolato tutto l'algoritmo per punti.

3.4.3 Schema dell'algoritmo per lo *sticky* HDP-HMM

Date le probabilità di transizione π, e la distribuzione dei parametri β e θ^{**} relativi all'iterazione $n-1$:

1. Inizializzando $m_{t+1,t}(k) = 1$, calcolare la sequenza dei messaggi all'indietro calcolando

   $$m_{t,t-1}(k) = \sum_{j=1}^{L} \pi_k(j) N(x_t; \mu_j, \Sigma_j) m_{t+1,t}(j)$$

   per ogni intero $k \in [1, L]$.

2. Campionare la sequenza degli stati, andando ora in avanti, iniziando cioè con $n_{jk} = 0 \; \forall j, k$ e calcolando:

   (a) la probabilità

   $$f_k(x_t) = \pi_{z_{t-1}}(k) N(x_t; \mu_k, \Sigma_k) m_{t+1,t}(k)$$

(b) l'assegnazione dello stato z_t

$$z_t \sim \sum_{k=1}^{L} f_k(x_t)\delta(z_t, k)$$

Aggiornare poi $n_{z_{t-1}z_t}$ e tutte le statistiche relative allo stato cui è stato assegnato x_t.

3. Una volta ottenuta tutta la sequenza di stati, ottenere il campionamento delle variabili ausiliarie come segue:

   (a) porre $n = 0$ e $m_{jk} = 0\ \forall j,k \in 1,\ldots,K$. Per ogni coppia (j,k) si effettua un'estrazione da una variabile aleatoria bernoulliana di parametro $\frac{\alpha\beta_k+\xi\delta(j,k)}{n+\alpha\beta_k+\xi\delta(j,k)}$. Si incrementa n e, se il risultato dell'estrazione è 1, si incrementa anche m_{jk}.

   (b) Per ogni j estrarre il numero di volte che la variabile di rivalutazione ω ha fatto cambiare il piatto scelto:

   $$\omega_{j\cdot} \sim Bin\left(m_{jj}, \frac{\rho}{\rho+\beta_j(1-\rho)}\right).$$

   (c) Infine porre

   $$\bar{m}_{jk} = \begin{cases} m_{jk} & j \neq k \\ m_{jj} - \omega_{j\cdot} & j = k \end{cases}$$

4. A questo punto si possono aggiornare le stime di

$$\begin{aligned} \beta &\sim Dir(\gamma/L + \bar{m}_1, \ldots, \gamma/L + \bar{m}_L) \\ \pi_k &\sim Dir(\alpha\beta_1 + n_{k1}, \ldots, \alpha\beta_k + \xi + n_{kk}, \ldots, \alpha\beta_L + n_{kL}) \\ \theta_k^{**} &\sim p(\theta_k^{**}|\mathrm{Y}_k) \end{aligned}$$

   dove X_k è l'insieme delle osservazioni assegnate allo stato θ_k^{**}.

5. Aggiornare le stime dei parametri α, γ, ξ.

3.4.4 Dettagli

Per poter completare l'esposizione dell'algoritmo mancano ancora un paio di dettagli. Questi sono stati lasciati per ultimi dato che non fanno parte del nucleo teorico riguardante lo *sticky* HDP-HMM, ma non possono essere trascurati.
Il primo di questi consiste nella specificazione di θ_k^{**}. Esso è il parametro che governa la funzione di distribuzione, pertanto nel nostro caso non è nient'altro che il vettore (μ_k, σ_k) formato da media e deviazione standard della distribuzione normale. Si può condurre inferenza Bayesiana su θ_k^{**} ponendo *a priori* $\mu_k \sim N(\mu_0, \sigma_0)$ e $\sigma_k \sim IW(\nu, \Delta)$, intendendo con IW la distribuzione di Wishart inversa.
Detto $|X_k|$ il numero di osservazioni assegnate a θ_k^{**}, le distribuzioni *full-conditional a posteriori* sono

$$\begin{aligned}\mu_k|\sigma_k &\sim N(\bar{\mu}_k, \bar{\sigma}_k)\\ \sigma_k|\mu_k &\sim IW(\bar{\nu}_k\bar{\Delta}_k, \bar{\nu}_k)\end{aligned}$$

con

$$\begin{aligned}\bar{\nu}_k &= \nu + |X_k|\\ \bar{\nu}_k\bar{\Delta}_k &= \nu\Delta + \sum_{t\in X_k}(x_t - \mu_k)^2\\ \bar{\sigma}_k &= (\sigma_0^{-1} + |X_k|\sigma_k^{-1})^{-1}\\ \bar{\mu}_k &= \bar{\sigma}_k\left(\sigma_0^{-1}\mu_0 + \sigma_k\sum_{t\in X_k}x_t\right).\end{aligned}$$

L'altro dettaglio che manca consiste nel ricavare le distribuzioni dei parametri α, γ e ξ. Più precisamente, ricordando la parametrizzazione suggerita al par 3.5, è sufficiente ricavare le distribuzioni di $\alpha + \xi, \gamma$ e ρ. Sui primi due parametri la distribuzione *a priori* scelta è una $Gamma(a, b)$, mentre per ρ è stata scelta una $Beta(c, d)$. Supponendo di avere all'iterazione corrente un numero complessivo di J ristoranti, $\bar{K}$ piatti unici considerati e che $\bar{m}_{\cdot\cdot}$ sia il numero totale di piatti considerati, si ha che

$$
\begin{aligned}
p\left(\alpha+\xi|\left\{m_{j\cdot}\right\}_{j=1}^{J},\left\{n_{j\cdot}\right\}_{j=1}^{J}\right) &\propto p(\alpha+\xi)p\left(\left\{m_{j\cdot}\right\}_{j=1}^{J}|\alpha+\xi,\left\{n_{j\cdot}\right\}_{j=1}^{J}\right) \\
&\propto p(\alpha+\xi)(\alpha+\xi)^{m_{\cdot\cdot}}\prod_{j}\frac{\Gamma(\alpha+\xi)}{\Gamma(\alpha+\xi+n_{j\cdot})} \\
p(\gamma|\bar{K},\bar{m}_{\cdot\cdot}) &\propto p(\gamma)\gamma^{\bar{K}-1}(\gamma+\bar{m}_{\cdot\cdot})\int_{0}^{1}\lambda^{\gamma}(1-\lambda)^{\bar{m}_{\cdot\cdot}-1}d\lambda \\
p(\xi|\omega) &\propto Beta\left(\sum_{j}\omega_{j\cdot}+c, m_{\cdot\cdot}\sum_{j}\omega_{j\cdot}+d\right)
\end{aligned}
$$

La derivazione e il campionamento da queste distribuzioni richiede dei passaggi matematici e statistici non banali che non vengono riportati e per i quali si rimanda ancora a Fox et al. (2008).

Capitolo 4

Analisi dei risultati

In questo capitolo verranno presentati i risultati relativi ai modelli presentati nei due capitoli precedenti.
Il capitolo è diviso in due parti. Nella prima si presenteranno i risultati del modello FDP (capitolo 2), si cercherà di trovare un raggruppamento delle funzioni di prezzo osservate per notare alcuni andamenti caratteristici di voli con caratteristiche comuni quali il tragitto e la compagnia aerea che offre il viaggio.
Nella seconda parte si mostrerà come si possono utilizzare i risultati relativi al modello *sticky* HDP-HMM presentato nel capitolo 3 per avere una previsione più precisa dell'andamento di ogni singola funzione di prezzo.
Le analisi basate sul primo modello sono state condotte utilizzando il programma *open source* R, mentre per quanto riguarda il secondo si è utilizzato MATLAB.

4.1 Risultati del modello FDP

In fase di applicazione dell'algoritmo, data la mancanza di informazioni *a priori*, sono stati scelti degli iperparametri che rendessero le distribuzioni *a priori* poco informative.

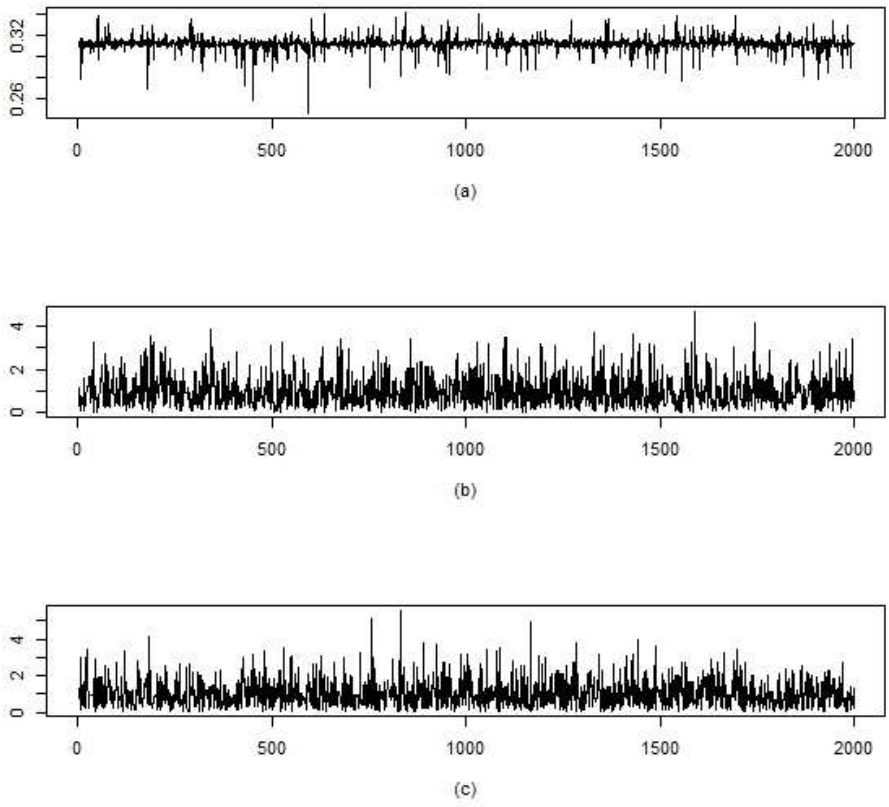

Figura 4.1: *Traceplot di τ (a) e dei parametri della funzione di autocovarianza (b)-(c).*

Il parametro α è stato fissato a 100, un valore molto alto tale da favorire la formazione di molti gruppi. La funzione media *a priori* del processo Gaussiano μ è stata fissata costante e uguale alla media delle osservazioni disponibili, mentre le altre distribuzioni *a priori* scelte sono:

$$\tau \sim Gamma(1,1),$$
$$\kappa_1 \sim Gamma(2,1),$$
$$\kappa_2 \sim Gamma(2,1).$$

L'algoritmo è stato eseguito per 2000 iterazioni e i *traceplot* dei parametri τ, κ_1 e κ_2 mostrati in figura sembrano confermare l'ipotesi di convergenza

della procedura.
La figura 4.1, inoltre, sembra mostrare l'assenza del periodo di *burn in*. In ogni caso, in tutte le analisi che seguiranno sono state escluse le prime 100 iterazioni dell'algoritmo.
Di seguito, vengono presentati i grafici e alcuni indici statistici delle densità a posteriori dei parametri.
Da essi è possibile notare come la distribuzione di τ sia molto concentrata e sostanzialmente simmetrica, mentre per κ_1 e κ_2 si hanno due densità molto simili tra loro.

	Media	Dev std	Primo quart.	Mediana	Terzo quart.
τ	0.312	0.007	0.311	0.313	0.315
κ_1	0.989	0.682	0.478	0.861	1.352
κ_2	0.995	0.716	0.449	0.817	1.380

Tabella 4.1: *Indici statistici delle distribuzioni a posteriori dei parametri.*

4.1.1 Costruzione dei cluster

I risultati dell'algoritmo consistono in 2000 diversi raggruppamenti dei dati. Questi raggruppamenti non sono coerenti tra loro, nel senso che a ogni iterazione viene generato un numero diverso di cluster. Inoltre, il gruppo numero 2 dell'iterazione i può essere completamente diverso da quello dell'iterazione $i + 1$. Questo fenomeno è detto *label switching* (Redner e Walker, 1984) e costituisce un problema di non poco conto quando l'analisi di raggruppamento è uno degli obiettivi preposti. Sono state proposte diverse soluzioni al problema, alcune molto raffinate come gli algoritmi di rietichettatura (Stephens, 1997). Qui si userà invece una procedura più semplice, suggerita da Medvedovic e Sivaganesan (2002).
Essa consiste nella costruzione di una matrice di distanze in cui l'elemento generico (i, j) è dato dal numero di volte che l'algoritmo ha allocato le corrispondenti osservazioni in due gruppi diversi.

Una volta ottenuta la matrice di distanze, si può applicare una qualsiasi procedura di *clustering* gerarchico agglomerativo (per una presentazione delle diverse modalità di raggruppamento si veda ad esempio Azzalini e Scarpa, 2011). In questo caso si è scelto il metodo del legame completo, ottenendo il seguente dendrogramma.

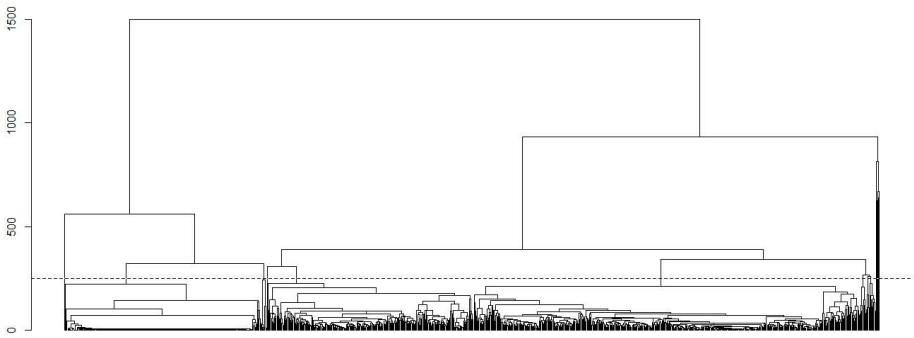

Figura 4.2: *Dendrogramma costruito con procedura agglomerativa utilizzando il metodo del legame completo.*

A questo punto si è scelto di tagliare il cluster ottenuto all'altezza indicata dalla linea tratteggiata. Si è così ottenuta una suddivisione in 30 gruppi, la maggior parte dei quali è formata da poche osservazioni molto diverse da tutte le altre.

4.1.2 Analisi dei gruppi

In figura 4.3 sono riportati gli andamenti osservati delle funzioni di prezzo appartenenti ai 6 gruppi costituiti da più unità. L'andamento delle funzioni va letto da destra verso sinistra, dato che nel dataset costruito P1 è il prezzo 24 ore prima della partenza. Anche basandosi esclusivamente su questi grafici, è possibile riconoscere alcuni tratti distintivi delle funzioni di prezzo presenti all'interno del gruppo. Si può ad esempio notare facilmente che nel

secondo gruppo vi sono funzioni di prezzo osservate relativamente corte, e che il prezzo sembra salire in modo esponenziale con l'avvicinarsi del giorno della partenza.

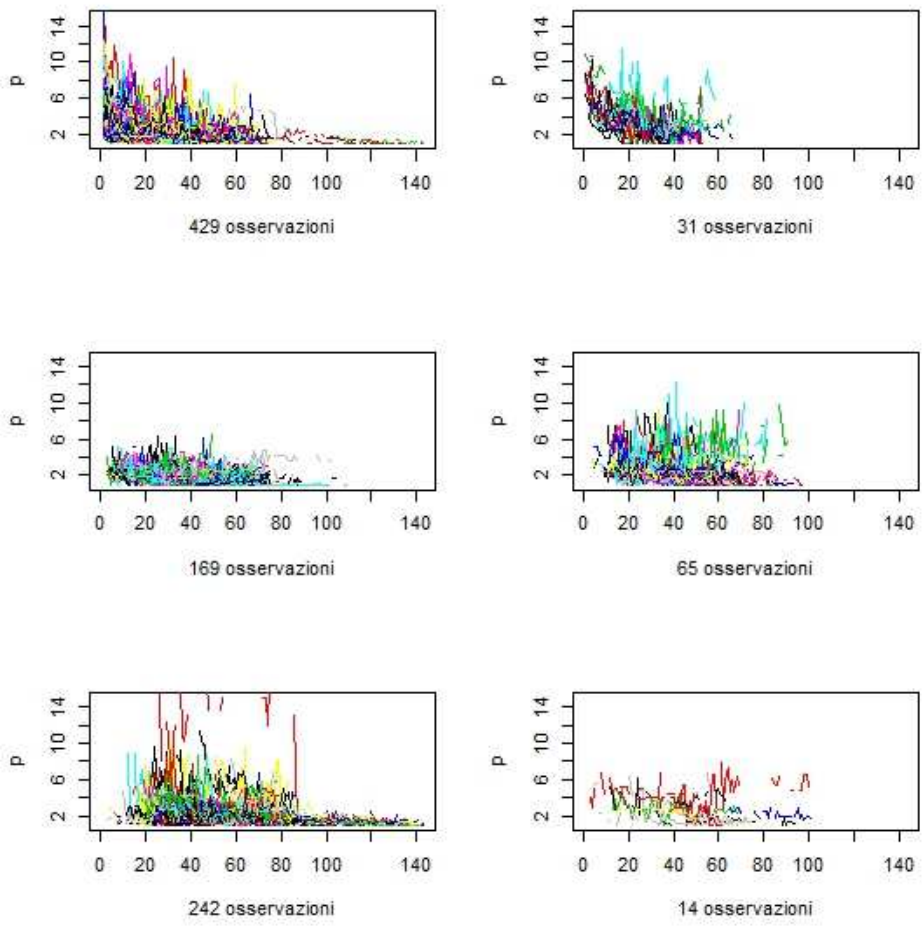

Figura 4.3: *Andamenti delle funzioni di prezzo suddivisi in gruppi. Sono qui rappresentati solo i gruppi contenenti più di 10 osservazioni. Si sono usati colori diversi per rappresentare meglio l'andamento generale nel tempo.*

Il gruppo 3 è invece costituito principalmente da voli aerei il cui prezzo aumenta in maniera minore rispetto al resto delle osservazioni. I gruppi 4 e 5 sono entrambi caratterizzati da funzioni di prezzo che terminano tutte alcu-

ni giorni prima del limite, probabilmente perché i posti sull'aereo vengono esauriti. La differenza tra i due gruppi consiste nella lunghezza e nella collocazione delle funzioni osservate, con il gruppo 5 che contiene quasi tutte le osservazioni con rilevazioni più lontane nel tempo.
Infine, è difficile analizzare i gruppi 1 e 6, il primo per la presenza di troppe osservazioni e il secondo per il motivo opposto.
Per avere un'idea migliore del tipo di funzioni di prezzo presenti in ognuno dei gruppi, la figura 4.4 mostra l'andamento medio in ognuno di essi.

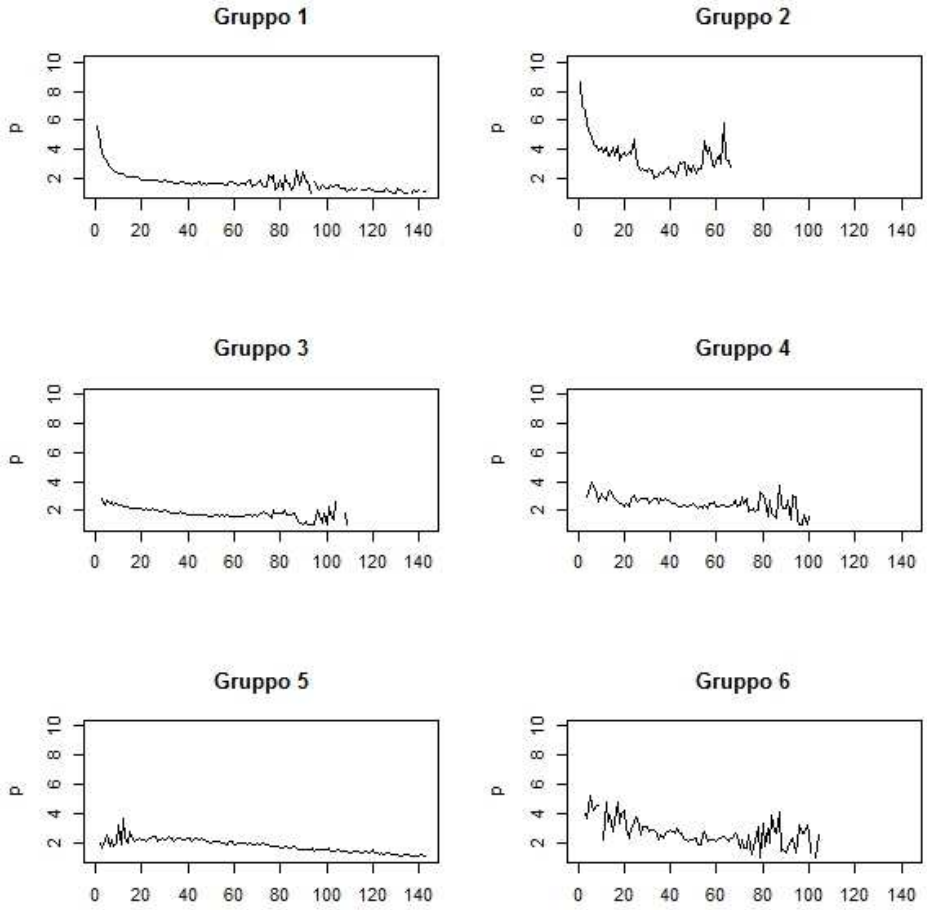

Figura 4.4: *Andamento medio osservato in ognuno dei cluster ottenuti.*

Si può notare ora che nel gruppo 1 si assiste a un andamento medio simile

a quello già descritto per il gruppo 2, con la differenza che qui vengono raggiunti valori di prezzo relativo minori. Inoltre, sembra che il gruppo 2 sia l'unico che non presenti un andamento monotono, dato che i valori più bassi dei prezzi rilevati non sono quelli più lontani dal giorno della partenza.
Dato che per ottenere questo raggruppamento dei dati non sono state considerate le variabili statiche, possiamo utilizzarle ora per cercare di descrivere meglio il tipo di gruppi che sono stati ottenuti.
Nelle tabelle seguenti, viene riassunta, gruppo per gruppo, la distribuzione delle variabili indicanti gli aeroporti di partenza e arrivo dei voli, il giorno della settimana in cui si parte e la compagnia aerea.

DEP AIRPORT G1	DEP DATE G1	ARR AIRPORT G1	ID AIRLINE G1
BGY :118	1: 21	BVA : 81	FR :337
BVA : 74	2: 29	FEZ : 31	U2 : 41
CIA : 38	3: 52	CMN : 26	IV : 17
MRS : 29	4: 75	BCN : 26	IG : 8
CDG : 27	5:133	STN : 25	W6 : 8
MXP : 24	6: 78	OPO : 17	TO : 5
(Other):119	7: 41	(Other):223	(Other): 13
DEP AIRPORT G2	DEP DATE G2	ARR AIRPORT G2	ID AIRLINE G2
BGY :17	1: 2	BVA :14	FR :31
BVA : 5	2: 0	BUD : 3	
CIA : 4	3: 5	CIA : 2	
BLQ : 1	4: 4	FUE : 2	
MRS : 1	5:10	STN : 2	
SUF : 1	6: 8	BGY : 1	
(Other): 2	7: 2	(Other): 7	

Tabella 4.2: *Riassunto della distribuzione delle variabili descrittive nei gruppi 1 e 2.*

La lettura di queste tabelle non è immediata, dato che bisogna tener conto della distribuzione generale delle variabili presentata alla fine del capitolo

DEP AIRPORT G3	DEP DATE G3	ARR AIRPORT G3	ID AIRLINE G3
BGY :45	1:14	BVA :26	FR :130
BVA :24	2: 9	RAK :19	U2 : 11
MRS :17	3:15	STN :12	IV : 7
BLQ :13	4:21	SUF :11	TO : 6
CIA :13	5:60	OPO :10	W6 : 4
MXP :10	6:32	PMO : 7	7M : 3
(Other):47	7:18	(Other):84	(Other): 8
DEP AIRPORT G4	DEP DATE G4	ARR AIRPORT G4	ID AIRLINE G4
BGY :23	1: 7	BVA :14	FR :53
BVA :16	2: 5	OPO : 7	U2 : 4
MRS : 6	3: 6	GRO : 6	IV : 2
CDG : 3	4:13	FEZ : 4	TO : 2
CIA : 2	5:14	CMN : 3	W6 : 2
FCO : 2	6:18	MRS : 3	7M : 2
(Other):13	7: 2	(Other):28	

Tabella 4.3: *Riassunto della distribuzione delle variabili descrittive nei gruppi 3 e 4*

1. Non deve stupire, infatti, che in tutti i gruppi gli aeroporti di partenza e arrivo più frequentati sono rispettivamente Orio al Serio (BGY) e Parigi Beauvais (BVA), o che la compagnia aerea più utilizzata sia costantemente Ryanair (FR), perché queste sono le modalità più comuni presenti nel dataset, e probabilmente anche nella totalità dei clienti che si rivolgono a Bravofly.
É comunque interessante notare che nel gruppo 2, queste caratteristiche si presentano in quasi la metà delle unità statistiche.
Inoltre, nei gruppi 4 e 5, la percentuale di voli in partenza di sabato è attorno al 25% ed è sensibilmente maggiore di quella osservata nei restanti gruppi ($\sim$ 18%).
Infine, si può notare che alcune destinazioni sono presenti in un solo gruppo, è questo il caso di Marrakech (RAK, gruppo 3).

DEP AIRPORT G5	DEP DATE G5	ARR AIRPORT G5	ID AIRLINE G5
BGY :45	1:23	STN : 38	FR :187
BVA :41	2:15	BVA : 37	U2 : 21
CIA :37	3:26	GRO : 20	IV : 14
BLQ :27	4:37	OPO : 13	IG : 6
MXP :25	5:68	BCN : 12	TO : 5
MRS :12	6:57	FEZ : 11	7M : 4
(Other):55	7:16	(Other):111	(Other): 5
DEP AIRPORT G6	DEP DATE G6	ARR AIRPORT G6	ID AIRLINE G6
BGY :5	1:4	OPO :3	FR :10
BVA :3	2:0	BVA :2	W6 : 2
FCO :2	3:2	PRG :2	0B : 1
EBU :2	4:1	MRS :1	7M : 1
NAP :1	5:3	ACE :1	
NTE :1	6:3	GRO :1	
	7:1	(Other):4	

Tabella 4.4: *Riassunto della distribuzione delle variabili descrittive nei gruppi 5 e 6.*

Con le informazioni in possesso si potrebbero ricercare altre informazioni riguardo alle osservazioni presenti nei gruppi. Ad esempio, si potrebbe usare l'ora di partenza per cercare di capire se alcuni di questi gruppi presentino maggiormente voli che partono di mattina o di sera, oppure confrontare gli aeroporti di partenza e arrivo per dividere i voli nazionali di uno Stato da quelli di un altro Stato e da quelli internazionali.
Tuttavia, ci si concentrerà ora sul metodo di previsione delle funzioni di prezzo dei viaggi aerei lasciando a futuri lavori il compito di individuare un *clustering* più accurato.
In figura 4.5 è stata disegnata la curva stimata per la funzione di prezzo già vista in figura 3.1. L'errore quadratico medio ottenuto per la curva è 0.3365.

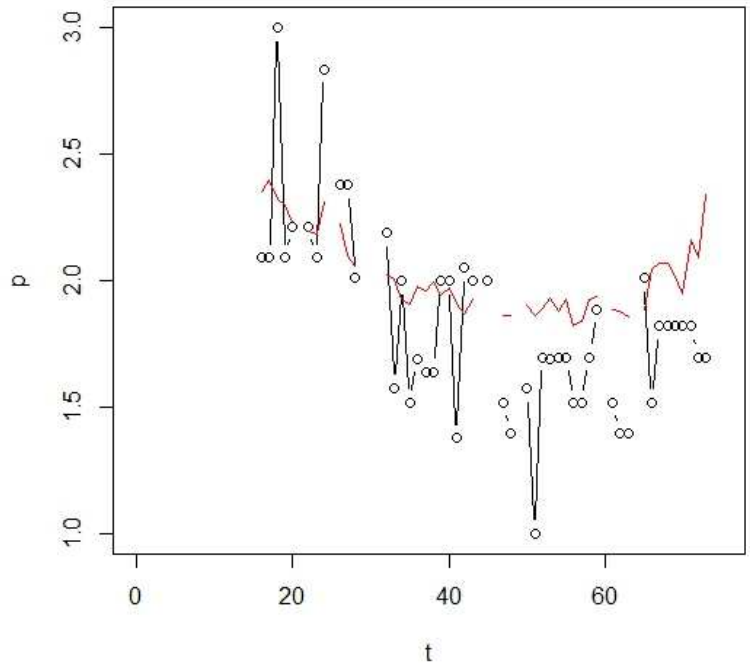

Figura 4.5: *Stime ottenute per la curva mostrata in figura 3.1 con il modello FDP.*

4.2 Risultati del modello *sticky* HDP-HMM

L'algoritmo presentato al capitolo 3, verrà applicato ad alcune specifiche funzioni di prezzo osservate. La prima di queste è la funzione mostrata in figura 3.1, che verrà usata per confrontare i due modelli in termini di precisione delle stime. In seguito, verranno stimate le catene di Markov sottostanti due funzioni di prezzo caratteristiche dei gruppi 2 e 3 ottenuti nel paragrafo 4.1.1. Per l'algoritmo presentato al capitolo 3, si sono scelte delle distribuzioni *a priori* che suggerissero la presenza di molti stati all'interno della catena di Markov, dato che in questa applicazione il prezzo osservato è direttamente collegato a uno stato della catena e perciò variazioni di esso devono allo stesso modo essere sintomo di una variazione di stato.
Per questo motivo il livello di approssimazione per l'HDP è stato posto a 20,

per la media degli stati si è usata una variabile aleatoria normale multivariata standard e come media della distribuzione di Wishart messa come *a priori* per la covarianza è stato scelto 0.01.
Le distribuzioni *a priori* scelte per gli altri parametri sono

$$\alpha \sim Gamma(1, 0.01),$$
$$\rho \sim Beta(1, 1),$$
$$\gamma \sim Gamma(1, 0.01).$$

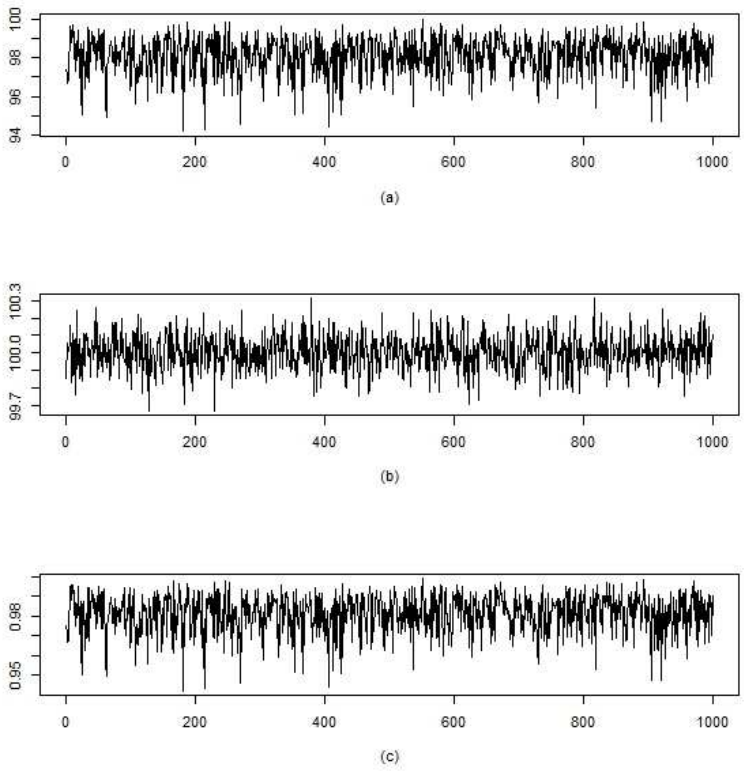

Figura 4.6: *Traceplot dei parametri del modello sticky HDP HMM. In (a) e (b) vi sono le stime di* α *e* γ *divise per* 10^6*, mentre in (c) le stime di* ρ*.*

Come per quanto fatto con il modello FDP, vengono proposti ora i *traceplot* di α, ρ e γ relativi all'algoritmo stimato sulla catena mostrata in figura 3.1.
Essi mostrano che la procedura è giunta a convergenza ed è pertanto possibile passare ad analizzare le catene di Markov stimate per questa e alcune altre funzioni di prezzo.
L'algoritmo ha effettuato 1000 iterazioni e, nonostante non ce ne fosse bisogno, è stato applicato ai dati riscalati per confrontare le stime con quelle del modello FDP.

4.2.1 Stime

Nella figura 4.7 è mostrata in nero la funzione di prezzo osservata e in rosso la stima risultante dal modello *sticky* HDP HMM.
Ci sono due aspetti da valutare. Il primo consiste nella capacità del modello di interpretare bene i cambiamenti di stato nella catena di Markov nascosta.
Il secondo, direttamente collegato al primo, è la capacità di previsione del prezzo ad un dato tempo t. A questo scopo, l'ordinata della curva che rappresenta la sequenza degli stati nella catena di Markov corrisponde alla media della funzione di emissione relativa allo stato da cui deriva la rilevazione secondo il modello stimato.
Dalla figura, si può notare come il modello catturi bene la persistenza dei prezzi nel tempo e come le stime siano molto precise, soprattutto per quanto riguarda la parte della curva più vicina al giorno di partenza.
Un problema che sembra non essere risolto riguarda alcuni picchi di prezzo presenti all'interno di periodi con prezzo costante. L'aver scelto un'*a priori* per le funzioni di emissione con poca variabilità, porta a interpretare questi discostamenti come degli effimeri e improbabili cambiamenti di stato nella catena di Markov. Se queste variazioni non dipendono invece dalla compagnia aerea, sarebbe opportuno che lo stato rimanesse costante. Una soluzione potrebbe consistere nel permettere alla funzione di emissione una maggiore variabilità, ma alcune prove sui dati hanno mostrato che ciò porterebbe a in-

terpretare la funzione di prezzo come risultante da una catena di Markov con un numero troppo basso di stati.

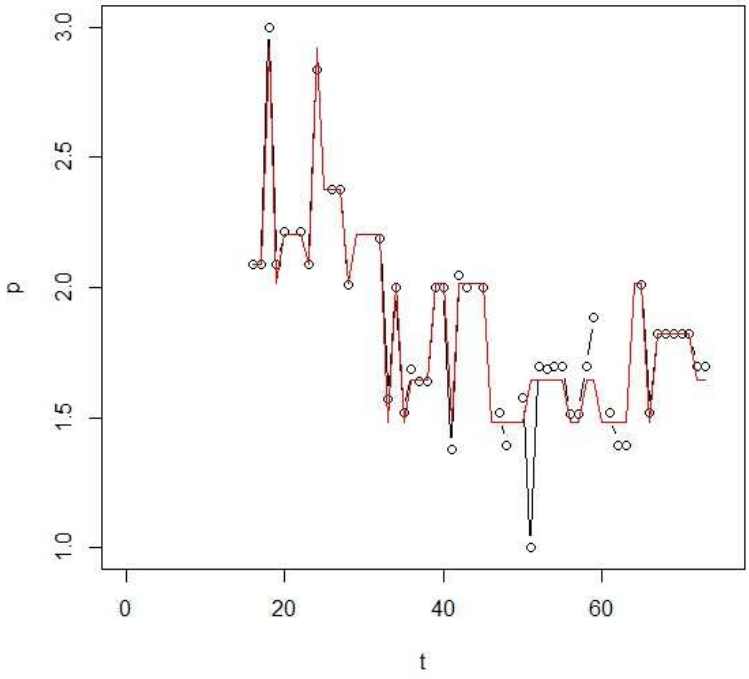

Figura 4.7: *Funzione di prezzo osservata e sequenza di stati stimata per i dati della figura 3.1. A ogni stato corrisponde la media della sua funzione di emissione.*

Com'era ampiamente prevedibile, l'errore quadratico medio è più basso e vale in questo caso 0.1093. Si è inoltre ottenuta una stima dell'errore tramite validazione incrociata *leave one out*, escludendo un dato alla volta dalla stima della sequenza degli stati (per approfondimenti si veda ad esempio, Azzalini e Scarpa, 2012). Essa vale 0.1657.
Infine, il modello è stato applicato a due funzioni di prezzo che sono rappresentative dei gruppi 2 e 3, dato che sono quelli che presentano degli andamenti maggiormente distintivi. In questo modo è possibile caratterizzare meglio le funzioni di prezzo presenti in questi due cluster. Le due funzioni

di prezzo scelte sono quelle che, all'interno del loro gruppo hanno la minor distanza media dall'andamento di prezzo medio mostrato in figura 4.4.

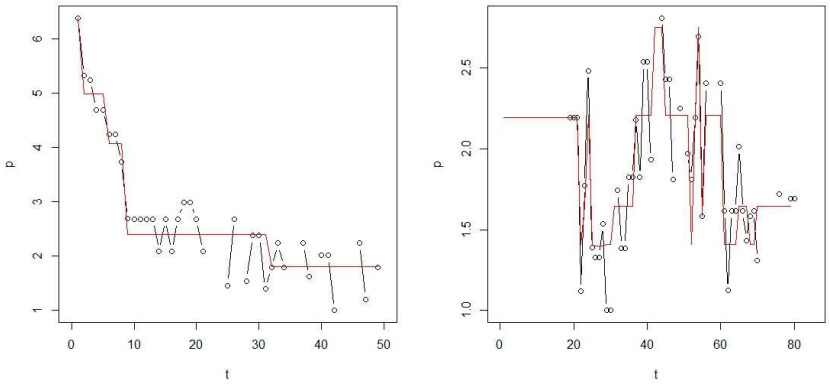

Figura 4.8: *Stime della sequenze di stati relative a due funzioni di prezzo rappresentative dei gruppi 2 (sinistra) e 3 (destra).*

Dal pannello di sinistra della figura 4.8 è possibile notare che, secondo la sequenza di stati stimata per il gruppo 2, fino a 10 giorni prima della partenza del volo, vi è un lungo periodo in cui i prezzi si mantengono sostanzialmente costanti. Terminato questo periodo, essi schizzano verso l'alto fino a più che quintuplicare il prezzo iniziale. Ricordando che il gruppo 2 era formato quasi esclusivamente da un tipo ben preciso di volo (Ryanair Bergamo-Parigi), questo modello suggerisce la possibilità di applicare il servizio di riserva del posto descritto a pagina 5 a questo tipo di voli, fino a 15 giorni prima della partenza.
Di diversa natura sono le considerazioni che si possono fare per il gruppo 3 guardando il pannello di destra. In questo caso, non ci sono lunghi periodi di persistenza e il modello *sticky* HDP HMM, non riesce a catturare bene la natura dell'andamento dei prezzi. Per questo motivo è difficile fare una previsione del prezzo di questo tipo di voli, che pur sono caratterizzati da un campo di variazione dei prezzi nel tempo più ristretto.

Conclusioni

In questa tesi si sono visti alcuni esempi di applicazioni della statistica Bayesiana non parametrica ai dati sui prezzi dei viaggi aerei forniti da Bravofly.
Con il modello basato sul FDP si è ottenuto un raggruppamento dei dati utilizzando esclusivamente l'informazione relativa all'andamento osservato della funzione. I risultati ottenuti non sono stati completamente soddisfacenti dato che alcuni gruppi presentavano un'elevata eterogeneità interna. Per ottenere dei cluster più omogenei, una strada potrebbe essere quella di utilizzare nel modello anche alcune variabili relative al volo considerato. Esse possono essere sia variabili statiche (ad esempio variabili qualitative relative alla nazione da cui parte il volo) sia dinamiche (ad esempio il numero di volte che il sito ha proposto il volo come soluzione di viaggio).
Lo *sticky* HDP-HMM è invece stato utilizzato per avere una previsione più precisa dei prezzi dei voli e per cercare di capire in quali giorni sia lecito attendersi un sensibile aumento (o, più raramente, una diminuzione) di prezzo. Il modello è stato applicato di volta in volta a una sola funzione di prezzo, ritenendo possibile l'estensione delle considerazioni fatte a tutte le osservazioni appartenenti a un certo gruppo (non necessariamente ricavato da un analisi di raggruppamento).
Un'analisi migliore potrebbe tuttavia essere condotta cercando di applicare questo modello direttamente a un insieme di funzioni, ad esempio tramite una gerarchia a più livelli, ottenendo allo stesso tempo un raggruppamento dei dati e un'accurata previsione dei livelli di prezzo.

Ringraziamenti

Dato che credo che questa sia la mia unica occasione per farlo, mi sento in dovere di ringraziare pubblicamente quanti hanno contribuito a farmi raggiungere questo traguardo.
Naturalmente, il grazie più speciale va rivolto ai miei genitori che hanno permesso che io facessi fino ad ora lo studente a tempo pieno e che mi hanno sempre sostenuto in tutte le mie scelte.
Un altro importante ringraziamento va a mio fratello, con cui non c'è neanche bisogno di parlare per intendersi (ma di tirarci qualche mazzata sì, sempre!).
C'è poi Elena, che da quando è entrata nella mia vita l'ha arricchita ogni giorno di sfumature nuove. A lei devo dire grazie soprattutto per non essersi mai accontentata di un semplice <<Va tutto bene!>> come risposta.
Questa tesi non sarebbe stata poi possibile senza l'aiuto di due Marco, che ironicamente, non sono legati solamente dall'omonimia. Entrambi hanno infatti contribuito attivamente a migliorare il mio rapporto con i software e sono per me, instancabile pigrone, un modello di impegno e professionalità. Senza dubbio i due più fighi che io conosca.
Non posso infine mancare di ringraziare tutti coloro che hanno reso divertenti i cinque anni passati all'Università, tra lezioni, ore di studio in gruppo intervallate da partite a carte (ma spesso era il contrario!), pranzi in mensa, serate in piazza, feste di Facoltà, feste e basta, vacanze... Forse ho imparato più da queste cose che dai professori!
Sperando di non dimenticare nessuno dico quindi grazie a Mattia, Tommaso, Andrea, Paolo, Omar, Giovanna, Sarah, Chiara, Massimiliano, Oriona,

Enrico, Giovanna (un'altra), Lorenzo, Ilaria, Laura, Elena (questa sì che è la stessa di prima invece!), Lara, Daniele, Davide. Sarebbe stato tutto molto più noioso e difficile se non vi avessi conosciuto.

Riferimenti bibliografici

Antoniak C. E. (1974). Mixtures of Dirichlet process with applications to Bayesian nonparametric problems. *The Annals of Statistic*, **2**, 1152-1174.

Azzalini A. e Scarpa B. (2011). *Data analysis and data mining*, Springer, Milano.

Blackwell D. e MacQueen J. (1973). Ferguson distributions via Polya urn schemes. *Tha Annals of Statistic*, **1**, 353-355.

Courty P. e Li H. (2000). Sequential screening. *Review of economic studies*, **67**, 697-717.

Cowans P. (2004). Information retrieval using hierarchical Dirichlet processes. *Proceedings of the annual international conference on research and development in information retrieval*, **27**, 564-565.

Dunson D. Nonparametric Bayes applications in biostatistics. In Hjort N. L., Holmes C., Muller P e Walker S. G. (2010). *Bayesian nonparametrics*, Cambridge University Press, Cambridge.

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

Fox E. B., Sudderth E. B., Jordan M. I. e Willsky A. S. (2008). An HDP-HMM for Systems with State Persistence. *Proceedings of the International Conference on Machine Learning*, Helsinki.

Fox E. B., Sudderth E. B., Jordan M. I. e Willsky A. S. (2011). A Sticky HDP-HMM with Application to Speaker Diarization *Annals of Applied Statistics*, Giugno 2011.

Goldwater S., Griffiths T. L. e Johnson M. (2006). Interpolating between types and tokens by estimating power-law generators. *Advances in neural information processing systems*, **18**.

Hjort N. L., Holmes C., Muller P e Walker S. G. (2010). *Bayesian nonparametrics*, Cambridge University Press, Cambridge.

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

MacEachern S. (1994). Estimating normal means with a conjugate style Dirichlet process priors. *Communication in statistics: simulation and computation*, **23**, 727-741.

Medvedovic, M. e Sivaganesan S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bionformatics*, **18**, 1194-1206.

Muller M. e Watanabe M. (2010). Advance purchase discounts versus clearance sales. *The economic Journal*, **120**, 1125-1148.

Netessine S. e Shumsky R. (2002). Introduction to the theory and practice of yield management. *Transactions on education*, **3**.

Nocke V. e Peitz M. (2008). Advance-purchase discount as a price discrimination device. CEPR discussion paper n° 6664.

Pitman J. (2002). Combinatorial stochastic processes. *Technical report 621*. Department of Statistics, University of California at Berkeley.

Rabiner L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceeding fo the IEEE*, **77**, 257-286.

Redner R. A. e Walker H. F. (1984). Mixture densities, maximum likelihood and EM algorithm. *SIAM review*, **26**, 195-239.

Robert C. P. e Casella G. (2010). *Introducing Monte Carlo methods with R*, Springer, Londra.

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

Stephens M. (1997). Bayesian methods for mixtures of normal distributions. *D. Phil. Thesis*, Department of Statistics, University of Oxford.

Teh Y. W., Jordan M. I., Beal M. J. e Blei D. M. (2006). Hierarchical Dirichlet processes. *Journal of the american statistical association*, **101**, 1566-1581.

Teh Y. W. e Jordan M. Hierarchical Bayesian nonparametric models with applications. In Hjort N. L., Holmes C., Muller P e Walker S. G. (2010). *Bayesian nonparametrics*, Cambridge University Press, Cambridge.

Wang S., Jank W. e Shmueli G. (2008). Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of business and economic statistics*, **26**, 144-160.