Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Tesi di Laurea

Le differenze delle percezioni olfattive nelle regioni italiane: un approccio bayesiano non parametrico alla fattorizzazione tensoriale

Relatore Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Correlatore Prof. Giancarlo Ottaviano
Dipartimento di Neuroscienze

Laureando Massimiliano Russo
Matricola N 1081034

Anno Accademico 2014/2015

Indice

Introduzione

L'olfatto è uno dei sensi più importanti nella vita quotidiana che consente di acquisire e far uso delle informazioni di un dato ambiente. Molti studi recenti suggeriscono che fattori ambientali influiscono in maniera importante su come gli odori vengano percepiti e su come vengano apprezzati dagli individui. L'Italia è una nazione in cui esistono ancora forti identità regionali, conseguenza delle diverse dominazioni susseguitesi nel corso dei secoli. Queste differenze sono evidenti, ad esempio, nella cucina in cui ancora oggi sono presenti piatti caratteristici di determinate aree. Questo lascia ipotizzare che anche nel giudizio associato agli odori possano esistere differenze considerevoli così come avviene in altre aree del mondo. Lo studio presentato in questa tesi tenta di mettere in evidenza tali differenze.
Da un punto di vista statistico il giudizio associato agli odori può essere visto come una variabile qualitativa ordinale. Nello studio proposto sono stati osservati 32 odori differenti in 4 aree geografiche.
L'interesse primario della tesi è, oltre che stabilire se esistano effettivamente delle differenze nel giudizio associato agli odori nelle diverse regioni italiane, presentare un modello che consenta di sfruttare la struttura di dipendenza che esiste tra le variabili trovando un compromesso tra la complessità del modello e la bontà della rappresentazione fornita. Trattando i dati come una tabella di contingenza essi possono essere considerati, da un punto di vista matematico, come un tensore. L'idea alla base del modello è, dato il tensore di partenza, approssimarlo tramite un tensore di rango inferiore con un'opportuna scomposizione. Questa rappresentazione di tipo matematico è equivalente ad un modello statistico che considera l'esistenza di classi latenti all'interno delle quali le variabili osservate sono tra loro condizionatamente indipendenti. In particolare, nei dati in esame l'attenzione è posta sulle differenze esistenti tra le diverse aree geografiche. La provenienza geografica è considerata, dunque, come una variabile esplicativa ed è di interesse stabilire come la misura di probabilità associata alla tabella di contingenza vari al variare di essa. Per inserire la dipendenza da una variabile qualitativa nella misura di probabilità si è ipotizzata l'esistenza di classi latenti, che la variabile esplicativa incida solo nella probabilità di appartenere ad una certa classe latente e che fissata quest'ultima non ci fossero differenze nel modo di percepire gli odori. L'ipotesi consiste nel supporre che all'interno della popolazione di riferimento esistano dei gruppi di individui che condividano lo stesso modo di giudicare gli odori. Nell'analisi considerata l'appartenenza ad uno dei gruppi non è di interesse primario ma rappresenta un modo di introdurre una semplificazione nel modello e di trattare la dipendenza esistente supponendo l'esistenza di una mistura di gruppi nella popolazione. Per la stima della misura di probabilità associata al

tensore si è utilizzato un approccio bayesiano non parametrico in modo da poter determinare direttamente dai dati il numero di gruppi necessario a descrivere adeguatamente il problema. Nel modello presentato si considera inoltre una distribuzione a-priori che inserisce un test di dipendenza dalla esplicativa direttamente nel modello. Questa distribuzione consente di generare dall'ipotesi di indipendenza o da quella di dipendenza calcolando la probabilità a-posteriori delle due ipotesi in modo da poter stabilire l'effetto di tale variabile sulla misura di probabilità.
Si sono proposte delle simulazioni per valutare le proprietà del modello sia sotto lo scenario di indipendenza, sia sotto quello di dipendenza in modo da valutarne le proprietà. Infine il modello è stato applicato ai dati disponibili ed i risultati ottenuti, in termini di previsione, sono stati confrontati con quelli di altri modello usati per problemi simili.

Capitolo 1

La percezione degli odori

In questo capitolo si considerano i dati disponibili sulla percezione degli odori nelle differenti regioni italiane. Si descrive come questi dati sono stati raccolti, si evidenziano alcune caratteristiche ed, infine, si effettuano alcuni test per mostrare la presenza di differenze significative tra le varie regioni.

1.1 Introduzione

L'olfatto sembra essere uno dei sensi più importanti nella vita di tutti i giorni. L'apprendimento conferisce flessibilità, consentendo agli individui di una data specie di acquisire e far uso in maniera appropriata delle informazioni in un dato ambiente. Un numero crescente di studi suggerisce che l'educazione gioca un ruolo fondamentale nella percezione degli odori (Schab, 1990). Si crede che questi stimoli contestuali siano codificati in episodi della memoria insieme agli eventi e che possano servire come innesco per il reperimento di dettagli dell'evento come le emozioni provate. Il potere degli odori di evocare memorie e associazioni e l'influenza delle esperienze sul giudizio sulla gradevolezza degli odori sono ben documentate in letteratura (per esempio Bensafi et al., 2007). In realtà, c'è una relazione molto stretta tra l'olfatto e le emozioni tale che nelle percezioni olfattive la risposta primaria di una persona alla percezione di un odore è stabilire se sia o meno gradevole. Questa conclusione è stata supportata da Bensafi et al. (2002) che hanno utilizzato metodi psicologici per studiare la possibile esistenza di un involontaria categorizzazione nell'olfatto. I loro risultati indicano che la percezione di odori spiacevoli provoca un aumento del battito cardiaco sia durante la percezione dell'odore sia nel momento in cui ne viene richiesto un giudizio. I risultati suggeriscono che gli individui involontariamente categorizzano gli odori a seconda della gradevolezza. Concentrandosi sul resoconto verbale delle emozioni indotte da una percezione olfattiva, in un recente studio, Croy et al. (2011) hanno concluso che solo un numero limitato di emozioni, ad esempio, felicità, ansia, disgusto, possono essere espresse verbalmente da uno stimolo olfattivo.

C'è un crescente riconoscimento nel fatto che le esperienze possano influenzare in maniera preponderante la percezione degli odori. In genere si riscontra una relazione positiva tra la gradevolezza degli odori e il giudizio su odori di ciò che è commestibile. È stato inoltre eviden-

ziato che tra diverse popolazioni (Giapponesi e Tedeschi) il giudizio relativo agli odori varia, suggerendo un effetto delle esperienze culturali specifiche (Ayabe-Kanamura et al., 1998). Inoltre, sembra che, fattori regionali influiscano sulla regolazione delle prestazioni olfattive, compresa la memoria olfattiva e la gradevolezza degli odori (Distel et al., 1999).
Oltre che per le variazioni genetiche, si presume che i fattori ambientali come l'istruzione e le esperienze giochino un ruolo importante nella regolazione delle prestazioni olfattive.
L'Italia racchiude una grande diversità ambientale nei suoi confini nazionali ed è altamente rinomata per una grande varietà di piatti regionali, vini e prodotti caseari derivanti dalla presenza contestuale di differenti ingredienti naturali e influenze culturali delle 20 regioni di cui è composta. È chiaro che, oltre alle contaminazioni nate dalle migrazioni interne e dalla presenza di una estesa cultura di massa, vi sono ancora delle grandi differenze tra le regioni riguardanti il gusto della popolazione per la cucina, la casa e i prodotti per l'igiene personale. Inoltre, si suppone l'esistenza di differenze nel patrimonio genetico tra le varie aree della penisola (Capocasa et al., 2013).

1.2 La rilevazione dei dati

Per evidenziare le differenze regionali (nord, centro, sud ed isole) i soggetti del campione sono stati reclutati a Padova per l'Italia del nord-est, a Roma per l'Italia centrale, a Napoli per l'Italia meridionale e in Sicilia per le isole. Tutti i soggetti sono stati studiati attraverso dei pennarelli profumati (*Sniffin'sticks*). Ogni soggetto considerato è stato sottoposto ad un pre-test: sono state proposti 16 differenti profumi e la possibilità di assegnare l'odore a 4 possibili classi. Ad ogni odore correttamente classificato è stato assegnato un punteggio e tutti i soggetti che hanno registrato un valore nell'intervallo costituito dalla media più o meno una volta la deviazione standard, differenziato per le classi di età come descritto in Hummel et al. (2007), sono stati considerati normosomici e quindi ammessi alla fase di test vera e propria. Nello studio sono stati reclutati 328 soggetti in salute di cui 191 femmine e 137 maschi. Il test olfattivo è stato condotto in una stanza tranquilla con adeguata ventilazione. Per la valutazione della gradevolezza degli odori, si è utilizzato un test sviluppato dal dipartimento di Chimica Analitica presso la facoltà di Tecnologia dell'università di Pardubice in Repubblica Ceca.
I pennarelli utilizzati per il test sono disponibili in commercio sotto forma di pennarelli con la punta di feltro, in cui, invece di utilizzare del colorante liquido il cilindro è stato riempito con degli odoranti.
In totale sono stati utilizzati 32 pennarelli riempiti con 2 mL di varie sostanze (Tabella 1.1). Come modalità, un pennarello aperto è stato posizionato a 2 cm da entrambe le narici e trattenuto per 4 secondi. Ai partecipanti è stato chiesto di posizionare il tono edonico relativo all'odore in una delle 4 categorie: piacevole (1), neutrale (2), spiacevole (3) e molto spiacevole (4). Dopo circa 15 secondi ai partecipanti è stato proposto l'odore successivo. Oltre al giudizio sugli odori somministrati, sono stati rilevati, il sesso e l'età dei soggetti, se il setto nasale fosse o meno indebolito e se il paziente avesse o meno carenze nella percezione olfattiva.

1.3 Divisione dei dati

Nell'ambito della valutazione dei modelli, spesso, per decidere quale sia il più opportuno, si applica una divisone dell'insieme di dati in modo da conservarne un parte su cui effettuare delle valutazioni. Questa tecnica viene utilizzata per trovare un modello che bilancia due entità in contrasto varianza e distorsione.
La componente di distorsione rappresenta di fatto la mancanza di conoscenza del meccanismo generatore dei dati. Se il meccanismo fosse noto si potrebbe scegliere il modello "corretto" per il problema e concentrarsi sulla varianza, in quanto la componente di distorsione sarebbe trascurabile.
L'insieme di dati utilizzato per la stima non è riutilizzabile per valutare il modello in quanto tenderebbe a premiare modelli che seguono fluttuazioni dei dati che derivano dall'errore di campionamento e che, non essendo strutturali, non sono propri del meccanismo generatore dei dati. L'uso di un insieme di verifica per la valutazione dei modelli previene tale problema, in quanto, sebbene affetti da nuova variabilità, consenteno di avere una valutazione del modello che tenga conto sia della varianza che della distorsione.
Idealmente la parte utilizzata per la valutazione dei modelli andrebbe tenuta separata da qualsiasi analisi ed utilizzata solo per le valutazioni finali dei modelli.
Per questo motivo e per non lasciarsi condizionare delle analisi iniziali, si è deciso di dividere l'insieme di dati disponibile, in due parti, fin dall'inizio e di utilizzare solo la parte dedicata alla stima per le prime considerazioni di tipo descrittivo presenti in questo capitolo.
Dei 328 soggetti disponibili solo 246, pari al 75% della numerosità campionaria, sono stati utilizzati nella fase descrittiva e di stima del modello. Il restante 25% è stato conservato per le valutazioni finali circa la capacità previsiva del modello.

numero del marcatore	odore/sostanza chimica	concentrazione originale	diluito	produttore/origine
1	rum	100	no	AROO s.r.o.
2	pineapple	100	no	AROO s.r.o.
3	fish composition	100	no	Aroma a.s.
4	buru babirusa	100	no	Aroma a.s.
5	propionic acid	100	water 1:25	faculty of chemical technology
6	almond	100	no	Dr.Oetker (Brasile)
7	n-butanol	100	water 1:25	faculty of chemical technology
8	formic acid	98	water 1:5	faculty of chemical technology
9	lemon	100	no	AROO s.r.o.
10	sour cherry	100	no	AROO s.r.o.
11	valeric acid	100	water 1:100	BASF
12	oleic acid	100	no	Chemapol
13	coconut	100	no	Kovandovi
14	water	100	no	faculty of chemical technology
15	vanilla	100	no	AROO s.r.o.
16	diesel fuel	100	no	OMV
17	valeraldehyde	97	water 1:125	faculty of chemical technology
18	women' parfume	100	no	Avon
19	octanoic acid	100	no	faculty of chemical technology
20	acetic acid	100	water 1:4	faculty of chemical technology
21	deer	100	no	Aroma a.s.
22	cyklohexanone	100	water 1:1	Apolda
23	propylene glycol	100	water 1:1	Germed
24	caproic acid	100	water 1:4	Reachim
25	men's parfume	100	no	NO II
26	n-butanol	100	water 1:5	faculty of chemical technology
27	fishing cat	100	no	Aroma a.s.
28	siberian musk deer	100	no	Aroma a.s.
29	strawberry	100	no	AROO s.r.o.
30	ethyl acetate	100	no	Penta
31	ethyl propionate	100	water 1:20	Lachema NP Brno
32	benzaldehyde	100	water 1:100	faculty of chemical technology

Tabella 1.1: lista marcatori utilizzati nell'analisi

1.4 Analisi Descrittive

I grafici successivi (Figure 1.1 e 1.2) mostrano le distribuzioni per diverse zone geografiche, dei giudizi sui 32 odori considerati.
Da una prima ispezione grafica (Figura 1.1 e 1.2) si può notare come esistano degli odori che sono percepiti in maniera molto simile. Si guardi ad esempio *lemon* che sembra essere per lo più gradevole a tutti indistintamente dalla zona geografica. Altri odori, invece, sembrano essere percepiti in maniera differente, come ad esempio, il *rum* che sembra piacere molto più al nord o *pineapple* che sembra essere gradito a tutti ma in particolare nelle isole. Può essere interessante vedere come il giudizio associato ad alcuni degli odori considerati possa variare a seconda dell'età, raggruppata in classi, e del sesso dei soggetti esaminati. Nella Tabella 1.2 sono riportate le frequenze relative divise per sesso e classe di età degli odori *rum* e *almond*. Le frequenze rappresentate sono normalizzate rispetto al sesso e alla classe di età considerata.

(a) *rum*

		<26	27-30	30-46	>46
Maschi	1	0.29	0.33	0.33	0.19
	2	0.27	0.60	0.50	0.43
	3	0.22	0.19	0.19	0.41
	4	0.04	0.00	0.00	0.00
Femmine	1	0.27	0.30	0.30	0.09
	2	0.56	0.41	0.41	0.44
	3	0.05	0.34	0.32	0.26
	4	0.00	0.00	0.00	0.00

(b) *almond*

		<26	27-30	30-46	>46
Maschi	1	0.33	0.43	0.38	0.10
	2	0.20	0.30	0.37	0.37
	3	0.15	0.44	0.26	0.41
	4	0.14	0.00	0.04	0.14
Femmine	1	0.27	0.42	0.33	0.09
	2	0.31	0.36	0.31	0.28
	3	0.26	0.26	0.34	0.29
	4	0.07	0.03	0.07	0.17

Tabella 1.2: Tabelle frequenze relative per sesso e classe di età

Dalla Tabella 1.2 si può notare come il *rum* piaccia per lo più agli uomini, in particolare ai più giovani, mentre un pò meno ai maschi con età superiore ai 46 anni. Quasi nessuno nel campione osservato ha reputato il *rum* molto sgradevole, solo il 4% dei ragazzi con età minore di 26 anni. Il *rum* non sembra essere particolarmente apprezzato dalle donne con età superiore ai 46 anni di cui solo il 9% contro il 19% dei maschi hanno espresso giudizio favorevole.
Per l'odore mandorla (*almond*) sembra piacere maggiormente sia agli uomini che alle donne in età compresa tra i 26 e i 46 anni. Tra gli uomini e le donne di età superiore ai 46 anni pochi hanno affermato che l'odore è piacevole (9% degli uomini e 10% delle donne), la maggior parte ha invece affermato che l'odore è neutro o spiacevole.

Figura 1.1: Grafici a barre delle frequenze relative osservate

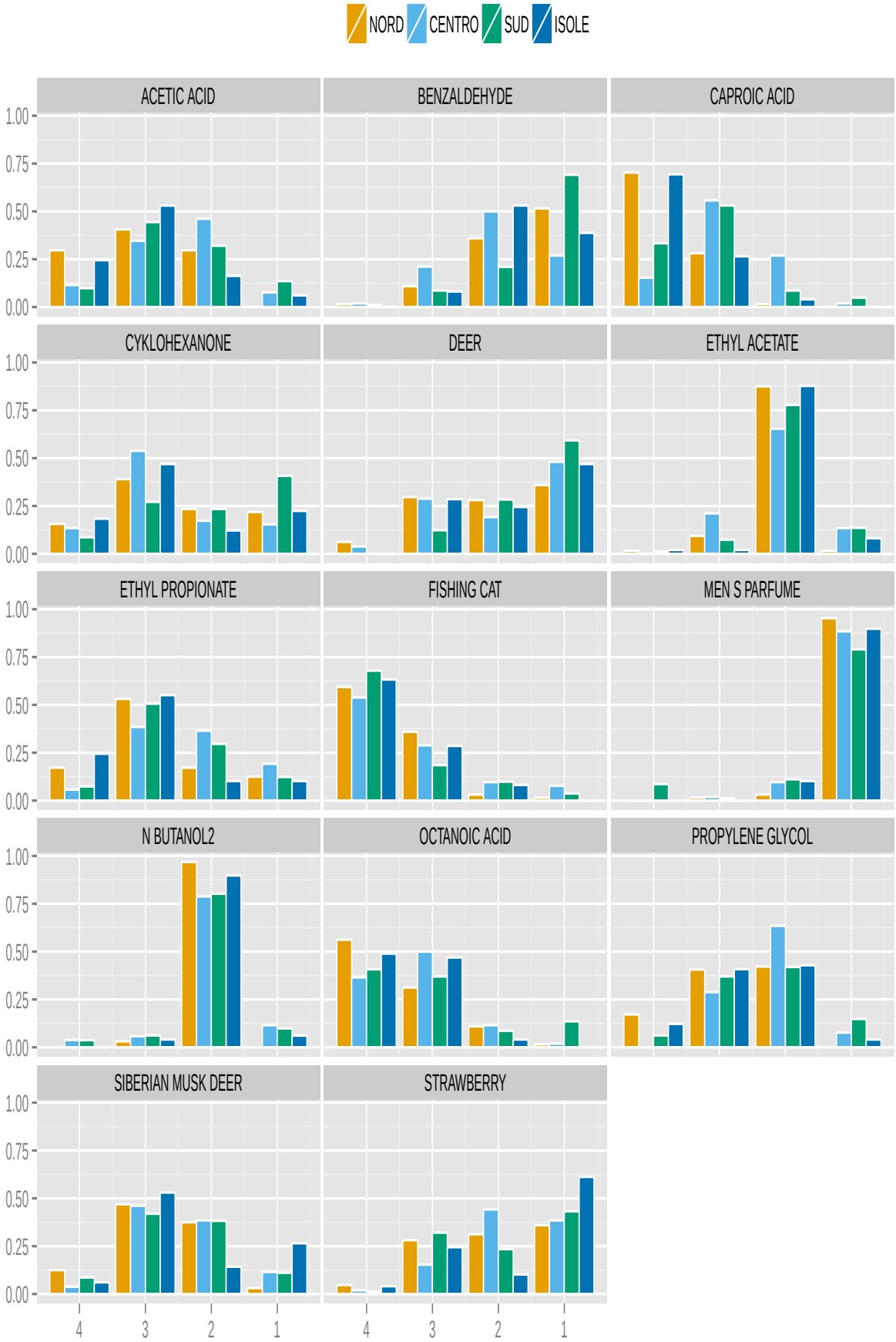


Figura 1.2: Grafici a barre delle frequenze relative osservate

1.5 Alcuni test sulle distribuzioni marginali

Considerando i 32 odori disponibili l'obiettivo principale è verificare se la loro percezione vari, o meno, in maniera statisticamente rilevante a seconda dell'area geografica.
Per verificare se, marginalmente, esista o meno tale differenza una possibilità è considerare se la variabile qualitativa relativa all'odore e la provenienza geografica siano o meno indipendenti. Inoltre, si può considerare il giudizio di gradevolezza come una variabile ordinale o semplicemente come una variabile qualitativa.
Questi test sono effettuati considerando singolarmente gli odori e non tenendo conto della struttura di dipendenza che esiste tra gli odori stessi.

1.5.1 Test di Pearson

Il test χ^2 di Pearson è uno dei test più applicati in presenza di variabili qualitative per valutare quanto sia plausibile che ogni differenza osservata tra insiemi di dati sia dovuta al caso. Le proprietà di questo test sono state studiate per la prima volta da Karl Pearson nel 1900.
L'ipotesi nulla è che la distribuzione di frequenza di un certo evento osservato nel campione sia coerente con una certa distribuzione teorica. Nel nostro caso consideriamo che, se gli odori fossero percepiti nello stesso modo nelle diverse aree geografiche le distribuzioni ottenute sarebbero tra loro indipendenti e, quindi, la probabilità di osservare una certa configurazione sarebbe data dal prodotto delle probabilità marginali.
Supponendo di avere due variabili categoriali $Y \in \{1, \dots, M\}$ e $X_j \in \{1, \dots, d_j\}$, dove nel nostro caso, Y rappresenta la provenienza geografica e X_j uno degli odori di cui è stato rilevato il giudizio di preferenza. La statistica test è data da

$$\chi^2 = \sum_{i=1}^{M} \sum_{j=1}^{d_j} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

Dove si è indicato con f_{ij} la frequenza congiunta delle modalità i e j per $i = 1, \dots, M$, $j = 1, \dots, d_j$ e con f_i ad f_j le rispettive marginali. La distribuzione asintotica della statistica test è data da un $\chi^2_{(M-1)(d_j-1)}$. Si e usato il pedice j per mantenere la notazione il più generale possibile anche se nel nostro $d_j = 4$ quasi per tutte le variabili e $M = 4$.
I risultati del test sono riportati in Tabella 1.3.

1.5.2 Logit ordinale

Una diversa possibilità è quella di considerare ognuno dei 32 odori disponibili come una variabile risposta ordinale. L'obiettivo è, dunque, quello di determinare se la probabilità, o il logit, di osservare un dato giudizio su un odore dipenda dalla provenienza geografica. Nel modello sono state inserite anche le altre variabili disponibili. Sfruttando il fatto che il giudizio sugli odori può essere visto come una misura ordinale, si può esprimere il logit della probabilità di osservare almeno un certo giudizio tramite la seguente funzione lineare (Azzalini

& Scarpa, 2012; Agresti, 2013)

$$\text{logit}(\mathbb{P}[X_j \leq k|z]) = \alpha_k + \beta^T z, \quad k = 1, \ldots, d_j$$

dove con il vettore z si è indicato sia la provenienza geografica che le altre variabili disponibili (sesso, età, setto nasale indebolito, carenze nella percezione olfattiva) di cui si è tenuto conto nell'effetture i test. Stimato il modello, per verificare se l'effetto della provenienza geografica fosse significativo, si è utilizzato il test rapporto di verosimiglianza. Si è calcolata la differenza di devianza tra il modello con provenienza geografica e quello senza, sfruttando l'approssimazione asintotica $\chi^2_{p-p_0}$, dove p rappresenta il numero di parametri nel modello stimato e p_0 il numero di parametri nel modello senza la provenienza geografica.
Questo test differisce dal test di Pearson del paragrafo precedente in quanto, inserendo l'ipotesi di linearita nel logit, tiene conto anche delle altre variabili misurate. Per ogni odore il test effettuato è dunque considerato al netto dei valori delle altre variabili. I risultati del test sono riportati in Tabella 1.3.

1.5.3 Test di Kruscal-Wallis e Test di Dunn

Il test di Kruskal-Wallis è uno dei test non parametrici più utilizzati per stabilire se una variabile continua, misurata al variare di un fattore, provenga o meno da un unica popolazione. Il test è basato sul confrontare la media dei ranghi e rappresenta l'alternativa non parametrica all'analisi della varianza ad una via.
Sebbene il test nasca per una variabile continua può essere riproposto in una versione per variabili ordinali sostituendo i ranghi con i *midranks*.
I *midranks* sono ottenuti assegnando ad ogni modalità la media del rango che sarebbe stato assegnato nel caso in cui non si fossero avute ripetizioni (vedere per esempio Agresti, 2010, pag. 201).

$$K = (n-1)\frac{\sum_{i=1}^{M}(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{M}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})}$$

Dove

- n è la numerosità totale del campione
- n_i è la numerosità totale del gruppo i
- r_{ij} rango osservazione j nel gruppo i
- $\bar{r}_{i\cdot} = \frac{\sum j=1^{n_i} r_{ij}}{n_i}$

la distribuzione asintotica del test è una χ^2_{M-1} gradi di libertà, dove nel nostro caso $M = 4$.
Stabilità l'esistenza di una differenza nell'effetto tra i vari gruppi è, inoltre, di interesse stabilire se la differenza interessa tutti i gruppi oppure solo alcuni di essi, per far ciò si può usare il test di Dunn (1964).
Il test di Dunn effettua un confronto per la dominanza stocastica facendo test a coppie

(a) Test di Pearson

	statistica	valore-p
propionic.acid	90.850	0.000
caproic.acid	82.565	0.000
almond	79.060	0.000
rum	62.620	0.000
valeraldehyde	45.184	0.000
benzaldehyde	40.224	0.000
pineapple	28.094	0.000
acetic.acid	33.746	0.000
fish.composition	30.578	0.000
valeric.acid	29.748	0.000
ethyl.acetate	29.487	0.001
octanoic.acid	28.978	0.001
propylene.glycol	28.270	0.001
n.butanol	24.903	0.003
water	23.318	0.006
formic.acid	23.035	0.006
deer	22.656	0.007
siberian.musk.deer	22.141	0.008
coconut	21.723	0.010
diesel.fuel	20.119	0.017
cyklohexanone	19.840	0.019
ethyl.propionate	19.618	0.020
sour.cherry	14.309	0.026
strawberry	18.293	0.032
men.s.parfume	18.093	0.034
oleic.acid	17.988	0.035
women..parfume	16.633	0.055
fishing.cat	15.241	0.085
n.butanol2	14.312	0.112
buru.babirusa	13.752	0.131
lemon	12.719	0.176
vanilla	2.778	0.836

(b) Logit ordinale

	statistica	valore-p
caproic.acid	76.920	0.000
propionic.acid	68.443	0.000
almond	43.183	0.000
valeraldehyde	37.674	0.000
rum	35.253	0.000
acetic.acid	27.447	0.000
benzaldehyde	25.434	0.000
propylene.glycol	18.636	0.000
valeric.acid	17.246	0.001
deer	17.123	0.001
pineapple	16.571	0.001
cyklohexanone	15.757	0.001
ethyl.propionate	14.859	0.002
n.butanol	14.696	0.002
sour.cherry	13.910	0.003
formic.acid	12.875	0.005
fish.composition	12.680	0.005
ethyl.acetate	11.132	0.011
buru.babirusa	10.112	0.018
women..parfume	9.463	0.024
men.s.parfume	8.798	0.032
water	7.805	0.050
strawberry	7.551	0.056
coconut	6.640	0.084
octanoic.acid	4.670	0.198
fishing.cat	4.614	0.202
siberian.musk.deer	3.937	0.268
lemon	3.445	0.328
diesel.fuel	3.058	0.383
vanilla	2.335	0.506
oleic.acid	1.554	0.670
n.butanol2	0.455	0.929

Tabella 1.3: Risultati test di Pearson e Logit ordinale ordinati per valori-p

seguendo il test di Kruskal-Wallis. L'interpretazione della dominanza stocastica richiede l'assunzione che la funzione di ripartizione di un gruppo non oltrepassi la funzione di ripartizione dell'altro. Quest'ipotesi nulla corrisponde a quella del test di Wilcoxon-Mann-Whitney sulla somma dei ranghi. Il testi di Dunn può essere interpretato come un test sulla differenza in mediana tra i vari gruppi. Il test effettua $M(M-1)/2$ (6 nel nostro caso) confronti a coppie. L'ipotesi nulla in ogni confronto è che la probabilità di osservare un valore casuale dal primo gruppo che sia maggiore rispetto a quella di osservare un valore casuale dal secondo gruppo è di un mezzo.
Visto che si stanno effettuando confronti multipli si è utilizzata, per il controllo del *family-wise error rate* la correzione proposta da Hochberg (1988). La correzione consiste nell'ordinare i valori-p dei singoli test dal più piccolo al più grande sostituendoli con $\max[1, \text{valore-}p \cdot i]$ dove i è l'indice dell'ordinamento.
Le statistiche test e i relativi valori-p sono riportati in Tabella 1.4.

1.6 Associazioni

Diversi tra gli odori analizzati sono comunemente utilizzati insieme in ricette o profumi, può quindi essere di interesse verificare se esiste un associazione tra di essi. Per valutare sia la direzione che il grado di associazione che esiste tra i differenti odori osservati si è fatto uso del coefficiente di Spearman (Agresti, 2013).
Il coefficiente di Spearman è una misura di associazione tra le variabili in cui l'unica ipotesi è che le variabili coinvolte nel calcolo siano almeno ordinabili, ovvero può essere utilizzato con variabili continue, quantitative discrete o qualitative ordinali. Indicando con r_i ed s_i i *midranks* relativi a due odori e con $\bar{r}_i$ ed $\bar{s}_i$ le loro medie il coefficiente di Spearman può essere calcolato come

$$\rho_s = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

Questo coefficiente stabilisce quanto bene la relazione tra due variabili può essere rappresentata da una funzione monotona. Il segno del coefficiente indica la direzione della relazione ed il suo campo di variazione è tra -1 e 1.
Considerando due variabili X e Y delle qualisi vuole calcolare l'associazione, un coefficiente di correlazione in valore assoluto indica una perfetta relazione monotona tra le variabili. Una relazione monotona crescente perfetta indica che per ogni coppia di valori x_i e y_i e x_j y_j, $x_i - x_j$ e $y_i - y_j$ hanno sempre lo stesso segno. Al contrario, una relazione monotona di tipo decrescente implica che le precedenti differenze hanno sempre segno opposto. Il coefficiente di correlazione di Spearman è spesso descritto come non parametrico. Tale affermazione può avere un duplice significato. In primo luogo, contrariamente alla correlazione di Pearson che misura solo l'associazione di tipo lineare la correlazione di Spearman misura una qualsiasi relazione monotona e la seconda è che la distribuzione campionaria esatta può essere ottenuta senza conoscere la distribuzione congiunta di X e Y. Esso può essere interpretato come un caso particolare del coefficiente di correlazione in cui i dati originali sono sostituiti dai ranghi.
I valori della statistica per tutte le coppie di odori sono riportati in Figura 1.3

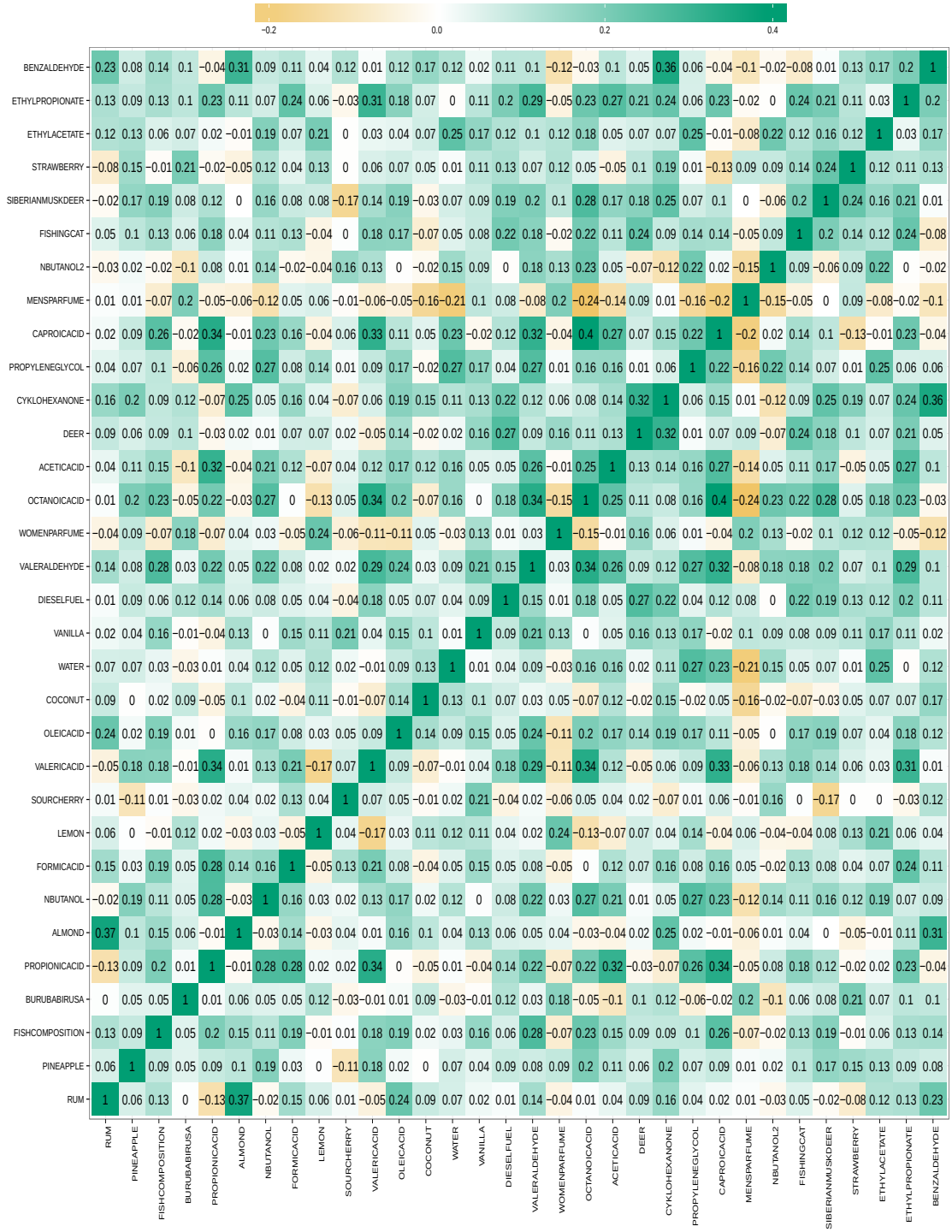


Figura 1.3: *heat-map* delle associazioni tra gli odori (correlazione di Spearman) dove il verde indica valori alti e il giallo valori bassi

	NORD-CENTRO	NORD-SUD	SUD-CENTRO	NORD-ISOLE	ISOLE-CENTRO	ISOLE-SUD
rum	5.194	3.006	-2.729	5.335	0.159	2.893
	0.000	**0.005**	**0.006**	**0.000**	0.437	**0.006**
pineapple	-1.777	-3.496	-1.504	-4.034	-2.241	-0.985
	0.113	**0.001**	0.133	**0.000**	**0.050**	0.162
fish composition	-4.055	-1.422	3.044	-1.048	2.962	0.250
	0.000	0.232	**0.006**	0.295	**0.006**	0.401
buru.babirusa	-0.280	1.391	1.680	-1.402	-1.111	-2.904
	0.390	0.246	0.232	0.322	0.266	**0.011**
propionic acid	-8.358	-4.288	4.934	-3.368	4.909	0.524
	0.000	**0.000**	**0.000**	**0.001**	**0.000**	0.300
almond	4.987	4.799	-0.732	6.006	1.027	1.866
	0.000	**0.000**	0.232	**0.000**	0.304	**0.093**
n butanol	-3.963	-1.947	2.424	-1.803	2.123	-0.062
	0.000	0.077	**0.038**	**0.071**	**0.067**	0.475
formic acid	-1.796	1.376	3.329	1.954	3.705	0.789
	0.109	0.169	**0.002**	0.101	**0.001**	0.215
lemon	-0.309	-0.891	-0.541	-1.420	-1.102	-0.681
	0.379	0.746	0.588	0.466	0.676	0.744
sour cherry	2.297	1.407	-1.131	3.540	1.239	2.498
	0.043	0.239	0.129	**0.001**	0.215	**0.031**
valeric acid	-3.818	-2.363	1.855	0.162	3.926	2.501
	0.000	**0.027**	**0.064**	0.436	**0.000**	**0.025**
oleic acid	0.147	-0.036	-0.197	0.784	0.631	0.895
	0.883	0.486	1.000	1.000	1.000	1.000
coconut	1.327	-1.747	-3.181	-0.310	-1.616	1.378
	0.185	0.201	**0.004**	0.378	0.212	0.252
water	0.021	-2.114	-2.110	-0.665	-0.678	1.350
	0.492	0.104	0.087	0.506	0.746	0.354
vanilla	-0.284	1.021	1.320	0.365	0.641	-0.604
	0.388	0.768	0.561	0.715	1.000	0.819
diesel fuel	-0.866	-0.808	0.152	-1.433	-0.564	-0.776
	0.966	0.838	0.440	0.455	0.572	0.656
valeraldehyde	-4.577	-0.722	4.308	0.823	5.329	1.612
	0.000	0.235	**0.000**	0.411	**0.000**	0.161
women's parfume	-0.525	0.647	1.214	-1.975	-1.438	-2.801
	0.300	0.518	0.337	0.121	0.301	**0.015**
octatonic acid	-2.790	-1.964	1.121	-0.569	2.189	1.307
	0.016	**0.099**	0.262	0.285	**0.072**	0.287
acetic acid	-3.575	-3.877	0.094	-0.171	3.357	3.624
	0.001	**0.000**	0.462	0.864	**0.001**	**0.001**
deer	-1.266	-3.432	-2.000	-0.233	1.018	3.119
	0.308	**0.002**	0.091	0.408	0.309	**0.005**
cyklohexanone	0.474	-2.244	-2.736	1.236	0.755	3.561
	0.318	**0.050**	**0.016**	0.325	0.450	**0.001**
propylene glycol	-3.533	-3.308	0.610	-0.526	2.964	2.676
	0.001	**0.002**	0.542	0.300	**0.006**	**0.011**
caproic acid	-7.425	-4.806	3.400	-0.058	7.266	4.661
	0.000	**0.000**	**0.001**	0.477	**0.000**	**0.000**
men's parfume	0.296	2.497	2.141	1.060	0.757	-1.293
	0.384	**0.038**	**0.081**	0.434	0.449	0.392
n butanol2	0.098	0.001	-0.106	-0.329	-0.422	-0.362
	0.922	0.499	1.000	1.000	1.000	1.000
fishing cat	-1.333	0.653	2.107	0.805	2.112	0.241
	0.365	0.514	**0.088**	0.631	0.104	0.405
siberian musk deer	-1.779	-1.495	0.475	-2.111	-0.335	-0.844
	0.188	0.270	0.635	0.104	0.369	0.598
strawberry	-1.980	-1.238	0.950	-2.111	-0.136	-1.097
	0.119	0.432	0.342	0.104	0.446	0.409
ethyl acetate	0.818	-2.135	-3.005	-1.465	-2.257	0.493
	0.413	**0.066**	**0.008**	0.214	**0.060**	0.311
ethyl propinate	-2.710	-1.469	1.522	0.899	3.563	2.430
	0.017	0.142	0.192	0.184	**0.001**	**0.030**
benzaldehyde	3.099	-1.773	-5.149	1.631	-1.441	3.531
	0.004	0.114	**0.000**	0.103	**0.075**	**0.001**

Tabella 1.4: Risultati del test di Dunn

1.7 Discussione

Dai test effettuati si può notare, anzitutto, che molti degli odori considerati presentano delle differenze significative per come sono percepiti nelle diverse regioni italiane. Considerando una soglia di accettazione, per il valore-p, di 5% il test χ^2 di Pearson considera differenti 26 dei 32 odori considerati nell'analisi, mentre il test basato sul logit ordinale è leggermente più conservativo e considera solo 22 dei 32 odori considerati come percepiti in maniera differente. Alcune di tali significatività potrebbero essere imputate ad altre variabili (età, sesso, etc.) in quanto il test di Pearson non considera l'effetto di queste ultime. I risultati dei due test sono, comunque, abbastanza concordi salvo qualche odore (ad esempio *fish composition* e *octanoic acid*) che cambia sensibilmente posizione in classifica.
Dalla tabella Tabella 1.3 si nota come le prime posizioni siano occupate dalle stesse variabili indicando che tali odori sono percepiti in maniera significativamente differente nelle diverse aree geografiche. In particolare le prime 7 posizioni sono occupate dalle stesse variabili a meno di qualche permutazione. Solo alcune variabili sono disposte in maniera differente, in particolare, *octatonic acid* che è considerato significativo dal test di Pearson non lo è per il logit ordinale passando dalla posizione 11 a 25, e *buru babirusa* che invece non è considerato significativo dal test di Pearson ma lo è dal logit ordinale. Dal test di Dunn (Tabella 1.4) si evince che solo alcuni odori sono percepiti in maniera differente solo in alcune tra le aree geografiche considerate. Ad esempio l'odore *rum* sembra essere differente tra tutte le possibili coppie di regioni tranne tra isole e centro. L'odore *buru babirusa* invece non sembra essere percepito in maniera diversa tra le regioni, supportando la conclusione ottenuta con il test di Pearson. Al contrario, *oleic acid* è considerato significativo, dal test di Pearson non sembra presentare nessuna differenza nel modo di essere percepito tra le diverse coppie di aree geografiche.
Dalla analisi effettuate, in conclusione è comunque possibile affermare che esiste una significativa differenza nella percezione olfattiva di molti odori tra le diverse regioni italiane, così come era ipotizzabile dagli studi condotti in altre aree del mondo.

Capitolo 2

Alcune nozioni sui tensori

In questo capitolo si introducono alcune definizioni e proprietà dei tensori usate nel successivo capitolo, per la presentazione del modello. Una tabella di contingenza, frequentemente utilizzata per l'analisi di variabili qualitative, è matematicamente esprimibile come un tensore. Per una più ampia trattazione dei tensori, ed in particolare alla loro decomposizione, si rimanda a Kolda & Bader (2009) e annessa bibliografia.

2.1 Definizione e notazione

Un tensore è un vettore multidimensionale, talvolta chiamato con il termine inglese *array*-multidimensionale. Per darne una definizione rigorosa è prima necessario definire cosa sia un prodotto tensoriale.
Il prodotto tensoriale, indicato in genere con $\otimes$ è una generalizzazione di un operatore bilineare e può essere applicato a molteplici oggetti, in particolare a vettori, matrici e spazi vettoriali. Nel caso di vettori il prodotto tensoriale è semplicemente il prodotto esterno tra i vettori. Se applicato a spazi vettoriali il prodotto tensoriale è un modo di creare un nuovo spazio vettoriale la cui dimensione è data dal prodotto degli spazi di partenza. Da un punto di vista matematico, quindi, un tensore a N-vie o N-dimensionale è un elemento del prodotto tensoriale di N spazi vettoriali ognuno dei quali con il proprio sistema di coordinate. Tale nozione di tensore è differente da quella utilizzata in fisica e ingegneria, che in matematica va invece sotto il nome di campo tensoriale. Un tensore di ordine uno è un vettore, un tensore di ordine due una matrice, mentre in generale un tensore di ordine 3 o più è detto tensore di ordine superiore. Un tensore di ordine 3 è caratterizzato da 3 indici e può essere rappresentato da un parallelepipedo (Figura 2.1).
L'ordine di un tensore è definito come il numero di dimensioni del tensore chiamate anche numero di mode o di vie. In alcuni ambiti, l'ordine di un tensore è anche definito rango. Il termine rango in tale contesto è, però, ambiguo in quanto spesso rappresenta, come si vedrà in seguito, qualcosa di differente.
Le fibre di un tensore (dall'inglese *fiber*) sono l'equivalente multidimensionale delle righe e delle colonne di una matrice. Una fibra è definita fissando tutti gli indici meno uno. In una matrice una colonna è una fibra di moda 1, una riga una fibra di moda 2. Un tensore di

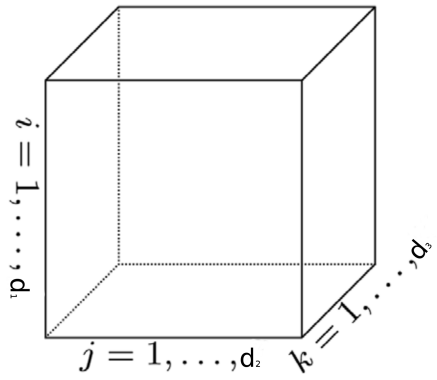
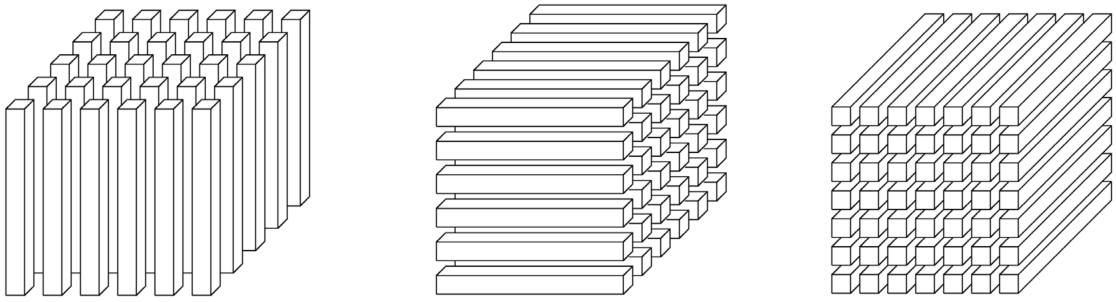
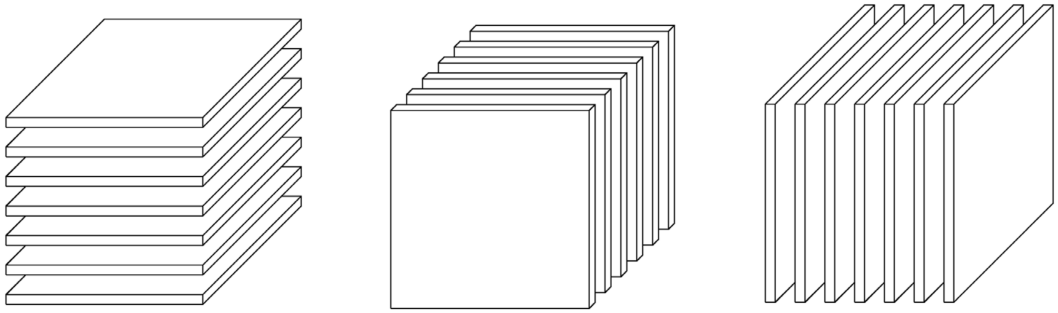


Figura 2.1: Tensore a 3 dimensioni

ordine 3 ha colonne, righe e tubi.
Gli *slice* di un tensore sono sezioni bidimensionali di un tensore, definiti fissando tutti gli indici tranne 2 (Figura 2.2).



(a) fibra di moda 1 (b) fibra di moda 2 (c) fibra di moda 3



(d) slice orizzontale (e) slice frontale (f) slice laterale

Figura 2.2: Fibre e slices di un tensore

Supponiamo di avere N vettori $\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(N)}$, un tensore N-dimensionale $\Psi \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ si dice di rango uno se può essere scritto come prodotto esterno di N vettori ad esempio

$$\Psi = \psi^{(1)} \otimes \psi^{(2)} \otimes \cdots \otimes \psi^{(N)}, \quad \psi^{(1)} \in \mathbb{R}^{d_1}, \ldots, \psi^{(2)} \in \mathbb{R}^{d_N}$$

il simbolo "$\otimes$" rappresenta il prodotto esterno ovvero ogni elemento del tensore è il prodotto

dei corrispondenti elementi dei vettori

$$\psi_{i_1,\dots,i_N} = \psi^{(1)}_{i1} \dots \psi^{(N)}_{i_N}$$

dove $\psi^{(1)}_{i_1}$ rappresenta l'i_1-esimo elemento del vettore $\psi^{(1)}$, e cosi via.

2.2 Rango e decomposizione di un tensore

Hitchcock (1927) introdusse l'idea della forma poliadica di un tensore (Figura 2.3), ovvero la possibilità di esprimere un tensore come somma del prodotto di un numero finito di tensori di rango 1. L'idea è quindi quella di poter scrivere un tensore $\Psi \in \mathbb{R}^{d_1 \times \dots \times d_N}$ nella forma

$$\Psi = \sum_{r=1}^{R} \psi^{(1)}_r \otimes \dots \otimes \psi^{(N)}_r, \quad \psi^{(1)}_r \in \mathbb{R}^{d_1}, \dots, \psi^{(N)}_r \in \mathbb{R}^{d_N}$$

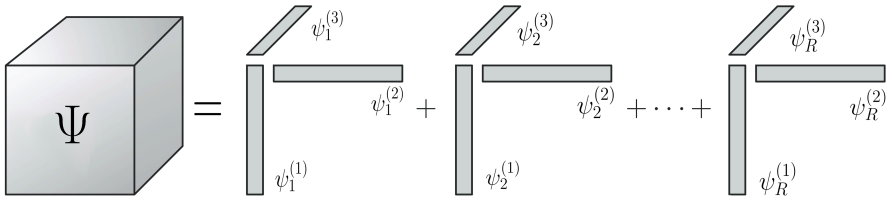


Figura 2.3: Forma poliadica di un tensore a 3 dimensioni

È spesso utile normalizzare la precedente espressione in modo che le colonne delle matrici risultanti dal prodotto abbiano una norma unitaria, per far ciò si può introdurre un vettore di pesi $\nu \in \mathbb{R}^R$ e riscrivere il precedente prodotto come

$$\Psi = \sum_{r=1}^{R} \nu_r \psi^{(1)}_r \otimes \dots \otimes \psi^{(N)}_r$$

Il rango di un tensore è definito come il più piccolo numero di tensori di rango uno che possono generare il tensore Ψ, ovvero l'indice R della formulazione precedente.
La definizione di rango di un tensore è l'analogo della definizione di rango di una matrice, ma le proprietà del rango di una matrice e di un tensore differiscono. Una delle principali differenze è che il rango reale di un tensore definito su $\mathbb{R}$ può differire da quello definito su $\mathbb{C}$. Un'altra differenza sostanziale è che non esiste un algoritmo semplice per il calcolo esatto del rango di un tensore il problema è infatti considerato *NP-hard*.
Nel contesto delle matrice, Eckart & Young (1936), hanno dimostrato che, volendo approssimare una matrice con una di rango inferiore, la migliore approssimazione di rango k di una matrice può essere ottenuta tramite i primi k fattori della scomposizione a valori singolari della matrice stessa. Questo tipo di risultato si generalizza ad un tensore di dimensioni superiori.
Considerando un $k < R$ è intuitivo che un tensore possa essere approssimato da una somma

di tensori di rango unitario, semplicemente troncando la precedente sommatoria al rango k.

$$\Psi = \sum_{i=1}^{k} \nu_r \psi_r^{(1)} \otimes \cdots \otimes \psi_r^{(N)}$$

In letteratura la formulazione precedente è nota come analisi parallela (*parallel analysis*) abbreviato in *PARAFAC* o equivalentemente analisi canonica (*canonical analysis*).
La *PARAFAC* è molto importante nell'ambito dei modelli statistici in quanto consente di ridurre la dimensionalità del problema in esame. Nell'ambito tensoriale, al contrario di quanto avviene con le matrici, non è detto che gli elementi della migliore approssimazione di rango k siano anche elementi della migliore approssimazione di rango $k+1$. Tale condizione rappresenta un problema nel momento in cui, per determinare il rango k per ottenere un'approssimazione adeguata si utilizza un algoritmo iterativo calcolando la *PARAFAC* per $k = 1, 2, \ldots$.
Questo implica che, se si vuole cercare un compromesso tra l'approssimazione effettuata e la dimensionalità del problema bisogna ricalcolare la *PARAFAC* ad ogni passo, per poi introdurre un criterio di scelta e decidere l'approssimazione opportuna al problema. Tale approccio può risultare computazionalmente oneroso. Fissato il numero di componenti esistono numerosi algoritmi per il calcolo della precedente scomposizione. Tra i vari algoritmi presenti uno dei più utilizzati è quello dei minimi quadrati alternati per i cui dettagli si rimanda al paper originale (Carroll & Chang, 1970).

Capitolo 3

Modello

Nell'analisi delle variabili qualitative, l'interesse principale è la struttura di dipendenza delle variabili considerate. In letteratura, sono stati presentati diversi metodi di analisi, di cui un ottima rassegna è Agresti (2013). Tra i differenti metodi proposti pochi sono, però, generalizzabili a contesti in cui la dimensione della tabella di contingenza è molto elevata. Considerando una tabella di contingenza come un tensore, la dimensione dello spazio associato è data dal prodotto del numero di categorie di ogni variabile osservata. Ogni singola cella del tensore rappresenta la frequenza di volte che si è verifica la data combinazione di modalità considerate. Al crescere del numero di variabili osservate, anche per un p modesto, il numero di celle della tabella di contingenza eccede in maniera significativa la numerosità campionaria (nel caso dei dati utilizzati il numero di celle è dell'ordine di 10^{20}) questo fa sì che i conteggi delle occorrenze di molte configurazioni siano nulli e, di conseguenza, il tensore molto sparso. L'obiettivo statistico è quello di ricostruire la distribuzione di probabilità congiunta delle variabili categoriali considerate.
Considerando una successione di variabili aleatorie qualitative $X_1, \ldots, X_p$ dove $X_j \in \{1, \ldots, d_j\}$, per $j = 1, \ldots, p$, la probabilità congiunta può essere espressa da

$$\pi = \mathbb{P}[X_1, \ldots, X_p] \qquad \text{dove} \qquad \pi \in \Pi_{(d_1,\ldots,d_p)}$$

dove $\Pi_{(d_1,\ldots,d_j)}$ è uno spazio tensoriale di dimensione $(d_1 \times \cdots \times d_p)$; inoltre gli elementi di π sono non negativi e la somma, rispetto a tutte le direzioni, è 1, equivalentemente π è una distribuzione di probabilità su $\Pi_{(d_1,\ldots,d_p)}$. Per ottenere una stima della distribuzione di probabilità precedente, dato che molti degli elementi del tensore sono vuoti, e data la dimensione elevata del numero di celle, è necessario ridurre la dimensionalità.
Come introdotto nel Capitolo 2, un modo per ridurre la dimensionalità è approssimare il tensore di interesse sfruttando la *PARAFAC*. Come puntualizzato in precedenza, tale rappresentazione pone il problema di decidere l'ordine di approssimazione del tensore prima di stimare un modello.
Il metodo bayesiano non parametrico proposto consente di non fissare il rango k a-priori ma di lasciare che siano i dati a stabilirlo in modo da ottenere un adeguata rappresentazione.

3.1 Un modello per la distribuzione congiunta di variabili qualitative

Dunson & Xing (2009) introducono un modello bayesiano non parametrico per la distribuzione congiunta di probabilità di una serie di variabili qualitative, rappresentate da un tensore.
Il punto di partenza è la scomposizione di un tensore tramite la *PARAFAC*.

$$\pi = \sum_{h=1}^{k} \nu_h \Psi_h \qquad \text{con } \Psi_h = \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)}$$

dove $\psi_h^{(1)}, \ldots, \psi_h^{(p)}$ sono dei vettori di probabilità che descrivono la probabilità di osservare una data modalità della variabile di interesse.
Oltre che da un punto di vista matematico, la precedente scomposizione può essere vista, da un punto di vista statistico, come un modello a variabili latenti.
Supponendo che ogni individuo sia associato ad una classe latente z_i e che, fissata la classe latente, la probabilità di osservare una certa modalità di una variabile sia condizionatamente indipendente dalle altre, la probabilità congiunta può essere scritta come

$$\begin{aligned}
\mathbb{P}[X_{i1} = c_1, \ldots, X_{ip} = c_p] &= \sum_{h=1}^{k} \mathbb{P}[X_{i1} = c_1, \ldots, X_{ip} = c_p | Z_i = h] \cdot \mathbb{P}[Zi = h] = \\
&= \sum_{h=1}^{k} \mathbb{P}[X_{i1} = c_1 | Z_i = h] \cdot \cdots \cdot \mathbb{P}[X_{ip} = c_p | Z_i = h] \cdot \mathbb{P}[Z_i = h] = \\
&= \sum_{h=1}^{k} \nu_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}
\end{aligned}$$

Dove ν_h indica la probabilità di appartenere alla classe latente h e con $\psi_{hc_j}^{(j)} = \mathbb{P}[X_{ij} = c_j | Z_i = h]$. Per un k fissato tale modello è noto come analisi delle strutture latenti (*latent structure analysis*) (Lazarsfeld et al., 1968).
Il modello consiste nel supporre che esista una variabile discreta non osservata con k modalità di cui le variabili osservate sono funzione. L'ipotesi di base è che, definita la classe latente di appartenenza, le variabili osservate siano condizionatamente indipendenti.
In altre parole, se fosse possibile creare una tabella di contingenza delle variabili osservate per ogni classe latente, le variabili osservate sarebbero indipendenti in ogni tabella. Senza tale assunzione, detta anche indipendenza locale, sarebbe necessario tener conto non solo dell'appartenenza ad una classe latente ma anche della dipendenza tra le variabili e non si avrebbe una riduzione di dimensionalità. L'ipotesi di indipendenza locale si riferisce soltanto alle classi latenti e non implica assolutamente che le variabili osservate siano tra di loro indipendenti in quanto esse sono una mistura di tutte le classi latenti.
L'analisi delle strutture latenti cerca, all'interno dei soggetti osservati, degli individui che presentano delle caratteristiche simili e li inserisce in uno stesso gruppo. Sebbene, come

detto in precedenza, per un k abbastanza grande la precedente specificazione consenta la rappresentazione di qualsiasi tipo di dipendenza tra i dati, il problema di come scegliere tale k rimane. In genere i dati sono molto sparsi e molte delle celle della tabella di contingenza considerate sono vuote, per cui, la stima di massima verosimiglianza dei parametri potrebbe non esistere anche per una scelta di k non eccessivamente elevata.
La scelta di k troppo piccolo, per ragioni di stima, potrebbe fornire un' approssimazione non adeguata della vera distribuzione multivariata e, quindi, l'inferenza sulla struttura di dipendenza delle osservazioni potrebbe essere distorta. Tale problema giustifica l'utilizzo dell'approccio bayesiano non parametrico di Dunson e Xing, che evita la scelta di un singolo k finito, consentendo al numero di gruppi latenti di crescere con la numerosità campionaria. Gli autori propongono di specificare per il tensore la seguente distribuzione a-priori.

$$\begin{aligned} \pi &= \sum_{h=1}^{\infty} \nu_h \Psi_h, \qquad \Psi_h = \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)} \\ \psi_h^{(j)} &\sim P_{0j} \text{ indipendentemente per } j = 1, \ldots, p, \quad h = 1, \ldots, \infty \\ \nu &\sim Q \end{aligned}$$

dove P_{0j} è una misura di probabilità sul simplesso d_j-dimensionale, e Q è una misura di probabilità sul simplesso infinito dimensionale. In particolare nell'articolo di Dunson e Xing si suggerisce di utilizzare una distribuzione di Dirichlet per P_{0j} ed un processo di Dirichlet, tramite la rappresentazione *stick-breaking*, per Q.
Considerando che la verosimiglianza, condizionatamente alle classe latente, è quella di una multinomiale e le ipotesi date per le distribuzioni a-priori il modello può essere riespresso nell'equivalente forma gerarchica

$$\begin{aligned} X_{ij}|Z_i = h &\sim \text{Multinom}\left\{(1, \ldots, d_j), \psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}\right\} \\ Z_i &\sim \sum_{h=1}^{\infty} \nu_h \delta_h, \qquad \nu_h = V_h \prod_{l<h} (1 - V_l), \qquad V_h \sim \text{Beta}(1, \alpha) \\ \psi_h^{(j)} &\sim \text{Dirichlet}(\alpha_{j1}, \ldots, \alpha_{jd_j}) \end{aligned}$$

Dove si è indicato con X_{ij} la j-esima variabile aleatoria discreta relativa all' i-esimo individuo; con l'espressione $\text{Multinom}\left\{(1, \ldots, d_j), \psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}\right\}$ la distribuzione di una variabile aleatoria discreta che può assumere valori $1, \ldots, d_j$ con probabilità $\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}$, elementi del vettore $\psi_h^{(j)}$.
Nel modello precedente le classi latenti sono scelte tramite la rappresentazione *stick-breaking* del processo di Dirichlet (Sethuraman, 1994).
Un campionamento da un processo di Dirichlet è composto dalla somma pesata di punti con massa di probabilità non nulla, tale rappresentazione conduce direttamente ad una formula costruttiva per il processo. La metafora per comprendere tale formulazione, nota come *stick-breaking* consiste nel considerare un asta di lunghezza unitaria e spezzarla in un

punto aleatorio V_1, assegnando tale probabilità alla prima componente della mistura, detta anche atomo. Il processo viene reiterato spezzando nuovamente la parte di asta rimanente e procedendo (Figura 3.1). Per costruzione la somma dei pesi è necessariamente unitaria. L'iperparametro α regola il numero di atomi presenti nel processo. Per valori di α prossimi

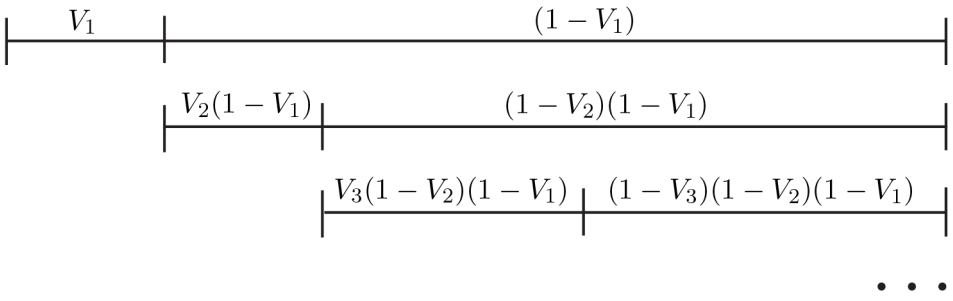


Figura 3.1: Rappresentazione grafica dello stick-breaking

allo 0 la massa di probabilità si concentra nel primo atomo, non si è quindi in presenza di una mistura, per valori di α piccoli sono presenti invece solo pochi atomi, viceversa al crescere del parametro il numero di atomi aumenta. Il parametro α è dunque un parametro cruciale in quanto determina il numero di componenti presenti; è quindi conveniente non fissare tale parametro in anticipo ma lasciare che siano i dati a stabilirlo tramite un iper-apriori (Escobar & West, 1995). Solitamente per mantenere la coniugazione e poter quindi sfruttare un *Gibbs Sampling* si utilizza una variabile aleatoria Gamma.
L'uso del processo di Dirichlet consente di non fissare in anticipo il numero di componenti della mistura ma di inferirlo dai dati. Il numero di componenti è potenzialmente infinito e nell'estrarre un campione di numerosità n, solo un numero finito di esse si realizza, ogni nuova unità può potenzialmente aggregarsi ad un gruppo esistente o crearne uno nuovo. Ishwaran & Zarepour (2002b) nel loro paper evidenziano come è possibile troncare opportunamente il processo di Dirichlet in quanto il numero di atomi è quasi certamente finito.

3.2 Label switching

Nel processo di Dirichelt, ad ogni iterazione, il numero di gruppi, così come la loro composizione può variare. Questa caratteristica che consente di determinare il numero opportuno di gruppi direttamente dai dati, genera per contro un fenomeno noto in letteratura come *label-switching* (Redner & Walker, 1984). Il *label-switching* consiste essenzialmente nel fatto che, sebbene gli individui vengano allocati insieme, ciò non avviene necessariamente nelle stesse classi ad ogni iterazione. La risoluzione di tale fenomeno è di particolare interesse quando il raggruppamento è l'obiettivo principale dell'analisi statistica, quando al contrario l'inferenza è marginale, ovvero ottenuta marginalizzando rispetto alle classi latenti, il *label-switching* non rappresenta un problema.
Nel caso preso in esame la creazione di gruppi è necessaria solo per semplificare la struttura di dipendenza e consentire una rappresentazione a dimensione ridotta della distribuzione di probabilità congiunta, per tal motivo non ci si è curati del problema.

3.3 Algoritmi di stima

Nell'articolo di Dunson e Xing è presentato un algoritmo per la stima del modello basato su una variazione dello *Slice Sampler* di Walker (2007). Questo algoritmo consente di non fissare un limite superiore per il numero di atomi nella rappresentazione *stick breaking* ma di determinare il numero, finito, di atomi con probabilità non nulla ad ogni iterazione. Sebbene l'algoritmo sia efficiente, può essere semplificato considerando i risultati studiati in Rousseau & Mengersen (2011) che suggeriscono, fissato un limite massimo sufficientemente elevato per le componenti di mistura, di generare la probabilità di appartenere ad una data componente della mistura da una Dirichlet. Tale procedura, con un opportuna distribuzione a-priori, favorisce la cancellazione automatica delle classi ridondanti come si è evidenziato anche dalle simulazioni (Capitolo 4). Inoltre, come notato in Ishwaran & Zarepour (2002a) rappresenta un approssimazione finita del Processo di Dirichlet di Ferguson.
Considerando tale semplificazione il modello di Dunson e Xing può essere ripresso nella forma gerarchica come

$$\begin{aligned} X_{ij}|Z_i = h &\sim \text{Multinom}\left\{(1,\dots,d_j), \psi_{h1}^{(j)},\dots,\psi_{hd_j}^{(j)}\right\} \\ Z_i &\sim \text{Multinom}\left\{(1,\dots,H), \nu_1,\dots,\nu_H\right\} \\ \nu_h &\sim \text{Dirichlet}\left(\frac{1}{H},\dots,\frac{1}{H}\right), \quad h = 1,\dots,H \\ \psi_h^{(j)} &\sim \text{Dirichlet}(\alpha_{j1},\dots,\alpha_{jd_j}) \end{aligned}$$

Questa nuova specificazione semplifica ulteriormente le procedure di stima, conducendo ad un algoritmo più efficiente per campionare dalla distribuzione a-posteriori.

3.4 Estensione del modello

Nei dati presentati nel Capitolo 1 la provenienza geografica gioca un ruolo importante in quanto è di principale interesse verificare se esistono variazioni nel giudizio degli odori al variare di essa.
Il modello proposto da Dunson e Xing non consente, però, di introdurre una dipendenza nella misura di probabilità da una variabile qualitativa come, nel nostro caso, la provenienza geografica. Se si è interessati alla probabilità che un certo individuo appartenga ad una certa classe del variabile qualitativa usata come predittore si può utilizzare il teorema di Bayes. Per il generico individuo i, si ha

$$\mathbb{P}[Y|X_{i1},\dots,X_{ip}] = \frac{p_Y(y)\mathbb{P}[X_{i1},\dots,X_{ip}|Y]}{\mathbb{P}[X_{i1},\dots,X_{ip}]}$$

dove la variabile aleatoria discreta $Y \in \{1,\dots,M\}$ rappresenta il predittore di interesse e $p_Y(y)$ la sua distribuzione di probabilità. Concentrandosi su $\mathbb{P}[X_{i1},\dots,X_{ip}|Y]$ essa potrebbe essere ottenuta semplicemente sfruttando il modello precedente e condizionandosi anche alla

Y, tale soluzione coincide, di fatto, con il simulare M (4 nel nostro caso) modelli differenti. Oltre che computazionalmente onerosa tale soluzione non è soddisfacente in quanto alcuni gruppi potrebbero essere costituiti da poche unità e, dato che i modelli sono stimati separatamente, non si effettuerebbe *borrowing* tra gli individui conducendo quindi ad un modello non efficiente.
Una soluzione differente consiste nel inserire la dipendenza dal predittore solo nei pesi della mistura esprimendo la probabilità condizionata come

$$\mathbb{P}[X_{i1}, \ldots, X_{ip}|Y] = \pi_y = \sum_{h=1}^{H} \nu_{hy} \Psi_h$$

Da un punto di vista statistico tale rappresentazione equivale a supporre che la probabilità che un individuo appartenga ad una certa classe latente dipenda dalla modalità osservata del predittore, mentre una volta stabilità la classe, la probabilità di osservare una data modalità di una covariata sia condizionatamente indipendente sia dai livelli del predittore sia dalle altre covariate. Tale approccio segue quanto già fatto nell'analisi delle strutture latenti, ovvero, lasciare che la dipendenza sia interamente assorbita dalle classi latenti. Il modello può quindi essere specificato nel seguente modo

$$\begin{aligned}
X_{ij}|Z_i = h &\sim \text{Multinom}\left\{(1, \ldots, d_j), \psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}\right\} \\
Z_i|y &\sim \text{Multinom}\left\{(1, \ldots, H), \nu_{1y}, \ldots, \nu_{Hy}\right\} \\
\nu_{hy} &\sim \text{Dirichlet}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right), \qquad \text{per } h = 1, \ldots, H \\
\psi_h^{(j)} &\sim \text{Dirichlet}(\alpha_{j1}, \ldots, \alpha_{jd_j})
\end{aligned}$$

Dalla precedente specificazione si è ottenuta la probabilità $\mathbb{P}[X_{i1}, \ldots, X_{ip}|Y]$ ma volendo fare inferenza su $\mathbb{P}[Y|X_{i1}, \ldots, X_{ip}]$ è necessario ottenere anche la probabilità di appartenere ad una certa categoria del predittore. Tale distribuzione può essere ottenuta da una multinomiale

$$Y \sim \text{Multinom}\left\{(1, \ldots, M), \varphi_1, \ldots, \varphi_M\right\}$$

dove le componenti del vettore di parametri $\varphi = (\varphi_1, \ldots, \varphi_M)$ indicano le probabilità di ottenere una delle M modalità della variabile. Per generare dalla distibuzione a-posteriori è possibile considerare la distribuzione a-priori coniugata al modello per il vettore di probabilità $\varphi = (\varphi_1, \ldots, \varphi_M)$ ovvero

$$\varphi \sim \text{Dirichlet}(\beta_1, \ldots, \beta_M)$$

3.5 Test Globale

Il modello precedente consente di stimare la classe di appartenenza di ogni individuo e può essere usato a fini predittivi. É, comunque, di interesse testare se esista o meno una differenza significativa tra i gruppi o se l'effetto del predittore sia statisticamente trascurabile. Tale ipotesi consiste nel verificare se la distribuzione congiunta del predittore e delle altre variabili sia fattorizzabile nel prodotto delle marginali o se almeno una delle condizionate sia diversa dalle altre

$$H_0 : \mathbb{P}[Y, X_1, \dots, X_p] = p_Y(y) \cdot \mathbb{P}[X_1, \dots, X_p] \quad \text{v.s.} \quad H_1 : \mathbb{P}[X_1, \dots, X_p|y] \neq \mathbb{P}[X_1, \dots, X_p|y'] \quad \text{per qualche } y, y'$$

Nel modello proposto, dato che la dipendenza dalla y è solo nei pesi della mistura, testare la differenza delle distribuzioni al variare della y equivale a testare

$$H_0 : \nu_{h1} = \cdots = \nu_{hM} \qquad \text{v.s.} \qquad H_1 : \nu_{hy} \neq \nu_{hy'} \qquad \text{per qualche } y, y'$$

Seguendo quanto proposto in Durante & Dunson (2014) questo test può essere inserito direttamente nel modello scegliendo un'opportuna distribuzione a-priori per i pesi della mistura. Definito il vettore dei pesi della mistura $\nu_y = (\nu_{y1}, \dots, \nu_{yH})$ si può il seguente modello gerarchico

$$\begin{aligned} \nu_y &= (1 - T)u + Tu_y \\ u &\sim \text{Dirichlet}\{\gamma_1, \dots, \gamma_H\}, \quad u_y \sim \text{Dirichlet}\{\gamma_1, \dots, \gamma_H\}, \quad y = 1, \dots, M \\ T &\sim \text{Ber}\{\mathbb{P}[H_1]\} \end{aligned}$$

T indica da quale ipotesi si sta generando, se $T = 1$ i pesi saranno generati in maniera indipendente da una Dirichlet al variare di y, mentre, se $T = 0$ non vi è differenza tra i gruppi e i pesi vengono da un'unica Dirichlet comune a tutte le y. L'uso di questa distribuzione a-priori non compromette il *Gibbs Sampling* in quanto si ottiene una forma chiusa per la *full conditional* di T. Indicando con $\text{B}(\gamma) = \prod_{h=1}^{H} \Gamma(\gamma_h)/\Gamma(\sum_{i=1}^{H} \gamma_i)$ la funzione Beta multivariata, costante di normalizzazione di una variabile di Dirichlet di parametri $\gamma = (\gamma_1, \dots, \gamma_H)$ e con

il vettore $\bar{n} = (\sum_{i=1}^{n} \mathbb{1}(Z_i = 1), \ldots, \sum_{i=1}^{n} \mathbb{1}(Z_i = H))$ si ha

$$\begin{aligned}
\mathbb{P}[T = 0|-] &\propto \mathbb{P}[H_0] \int \left\{ \prod_{i=1}^{n} \mathbb{P}[Z_i|u] \right\} \prod_{h=1}^{H} \frac{u_h^{\gamma_h - 1}}{\mathrm{B}(\gamma)} du = \\
&= \frac{\mathbb{P}[H_0]}{\mathrm{B}(\gamma)} \int \left\{ \prod_{i=1}^{n} \prod_{h=1}^{H} u_h^{\mathbb{1}(Z_i = h)} \right\} \prod_{h=1}^{H} u_h^{\gamma_h - 1} du = \\
&= \frac{\mathbb{P}[H_0]}{\mathrm{B}(\gamma)} \int \left\{ \prod_{h=1}^{H} u_h^{\sum_{i=1}^{n} \mathbb{1}(Z_i = h)} \right\} \prod_{h=1}^{H} u_h^{\gamma_h - 1} du = \\
&= \frac{\mathbb{P}[H_0] \cdot \mathrm{B}(\gamma + n_h)}{\mathrm{B}(\gamma)} \underbrace{\int \frac{\prod_{h=1}^{H} u_h^{\sum_{i=1}^{n} \mathbb{1}(Z_i = h) + \gamma_h - 1}}{\mathrm{B}(\gamma + n_h)} du}_{=1} = \\
&= \frac{\mathbb{P}[H_0] \cdot \mathrm{B}(\gamma + n_h)}{\mathrm{B}(\gamma)}
\end{aligned}$$

Con analoghi passaggi si può ottenere la *full conditional* di $T = 1$

$$\mathbb{P}[T = 1|-] \propto \frac{\mathbb{P}[H_1] \cdot \prod_{y=1}^{M} \mathrm{B}(\gamma + \bar{n}_{hy})}{\mathrm{B}(\gamma)}$$

dove $\bar{n}_y = (\sum_{i:y=y} \mathbb{1}(Z_i = 1), \ldots, \sum_{i:y=y} \mathbb{1}(Z_i = H))$
Normalizzando si ottiene che

$$\mathbb{P}[T = 1|-] = \frac{\mathbb{P}[H_1] \cdot \prod_{y=1}^{M} \mathrm{B}(\gamma + \bar{n}_y)/\mathrm{B}(\gamma)}{\mathbb{P}[H_0] \cdot \mathrm{B}(\gamma + \bar{n})/\mathrm{B}(\gamma) + \mathbb{P}[H_1] \cdot \prod_{y=1}^{M} \mathrm{B}(\gamma + \bar{n}_y)/\mathrm{B}(\gamma)}$$

per cui per stabilire da quale ipotesi generare basta estrarre T da una Bernoulliana con probabilità di successo $\mathbb{P}[T = 1|-]$.
Ricapitolando il modello, compreso di quest'ultimo passaggio, può essere espresso in forma gerarchica come

$$\begin{aligned}
Y &\sim \mathrm{Multinom}\left\{ (1, \ldots, M), \varphi_1, \ldots, \varphi_m \right\} \\
\varphi &\sim \mathrm{Dirichlet}(\beta_1, \ldots, \beta_M) \\
X_{ij}|Z_i = h &\sim \mathrm{Multinom}\left\{ (1, \ldots, d_j), \psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)} \right\} \\
Z_i|y &\sim \mathrm{Multinom}\left\{ (1, \ldots, H), \nu_{1y}, \ldots, \nu_{Hy} \right\} \\
\nu_y &= Tu + (1 - T)u_y \\
u &\sim \mathrm{Dirichlet}\{\gamma_1, \ldots, \gamma_H\}, \quad u_y \sim \mathrm{Dirichlet}\{\gamma_1, \ldots, \gamma_H\}, \quad y = 1, \ldots, M \\
T &\sim \mathrm{Ber}\{\mathbb{P}[H_1]\} \\
\psi_h^{(j)} &\sim \mathrm{Dirichlet}(\alpha_{j1}, \ldots, \alpha_{jd_j})
\end{aligned}$$

Campionando dal precedente modello si può quindi decidere in favore o meno dell'ipotesi nulla considerando la probabilità a-posteriori $\mathbb{P}[H_1|y, X]$ o il fattore di Bayes

$$\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \cdot \frac{\mathbb{P}[H_1|y, X]}{\mathbb{P}[H_0|y, X]}$$

entrambe queste quantità possono essere calcolate direttamente dall'*output* del *Gibbs Sampling*

3.6 Gibbs Sampling

Dalla formulazione gerarchica del modello, scritta nel paragrafo precendente, si ha che la distribuzione a-posteriori congiunta di interesse è data da

$$\mathbb{P}[\Psi, \nu, \varphi|X_1, \ldots, X_p, Y] = \mathbb{P}[Y|\varphi] \cdot \mathbb{P}[X_1, \ldots, X_p, Y|\Psi, \nu] \cdot \pi(\varphi) \cdot \pi(\Psi) \cdot \pi(\nu)$$

La distribuzione di $\mathbb{P}[Y|\varphi] \cdot \pi(\varphi)$ non dipende dalle da $X_1, \ldots, X_p$ ed è quella di un modello Dirichlet-Multinomiale.
Per quanto riguarda la restante parte essa può essere scritta come

$$\mathbb{P}[\Psi, \nu|X_1, \ldots, X_p, Y] \propto \sum_{h=1}^{H} \nu_{hy} \prod_{j=1}^{p} \prod_{c=1}^{d_j} [\psi_{hc}^{(j)}]^{\sum_{i=1}^{n} \mathbb{1}(X_{ij}=c)} \pi(\Psi)\pi(\nu)$$

Per questa distribuzione non è possibile ottenere la a posteriori in maniera analitica per il modello proposto. È, però, possibile, avendo specificato distribuzioni a-priori tutte coniugate, ottenere un efficiente *Gibbs Sampling* seguendo i seguenti passi e iterando fino a convergenza

1) Si estraggono le probabilità di osservare una data modalità per ognuna delle variabili $\psi_h^{(j)}$ per $j = 1, \ldots, p$ e $h = 1, \ldots, H$ dalla *full conditional*

$$\psi_h^{(j)}|- \sim \text{Dirichlet}\left\{\alpha_{j1} + \sum_{i:z_i=h} \mathbb{1}(X_{ij} = 1), \ldots, \alpha_{jd_j} + \sum_{i:z_i=h} \mathbb{1}(X_{ij} = d_j)\right\}$$

2) Per ogni individuo si identifica la classe latente di appartenenza, generando dalla variabile aleatoria multinomiale con probabilità

$$\mathbb{P}[z_i = k|-] = \frac{\nu_k \prod_{j=1}^{p} \psi_{kx_{ij}}^{(j)}}{\sum_{h=1}^{H} \nu_h \prod_{j=1}^{p} \psi_{hx_{ij}}^{(j)}}$$

3) L'indicatore dell'ipotesi dalla quale si sta genereando T è estratto da una Bernoulliana con probabilità di successo

$$\mathbb{P}[T = 1|-] = \frac{\mathbb{P}[H_1] \cdot \prod_{y=1}^{M} \text{B}(\gamma + \bar{n}_y)/\text{B}(\gamma)}{\mathbb{P}[H_0] \cdot \text{B}(\gamma + \bar{n})/\text{B}(\gamma) + \mathbb{P}[H_1] \cdot \prod_{y=1}^{M} \text{B}(\gamma + \bar{n}_y)/\text{B}(\gamma)}$$

4) Se $T = 0$ si pone $\nu_y = u$, $y = 1, \ldots, M$ e si aggiorna u dalla *full conditional*

$$\text{Dirichlet}\left\{\gamma_1 + \sum_{i=n} \mathbb{1}(z_i = 1), \ldots, \gamma_H + \sum_{i=n} \mathbb{1}(z_i = H)\right\}$$

Se invece $T = 1$ $\nu_y = u_y$, $y = 1, \ldots, M$ ognuno dei quali aggiornato indipendentemente da

$$\text{Dirichlet}\left\{\gamma_1 + \sum_{i:y_i=y} \mathbb{1}(z_i = 1), \ldots, \gamma_H + \sum_{i:y_i=y} \mathbb{1}(z_i = H)\right\}$$

5) Generare Y da

$$Y|- \sim \text{Dirichlet}\left\{\beta_1 + \sum_{i=1}^{n} \mathbb{1}(y_i = 1), \ldots, \beta_M + \sum_{i=1}^{n} \mathbb{1}(y_i = M)\right\}$$

3.7 Test locale

Sebbene il precedente test, incluso nel modello, consenta di affermare se esista o meno una differenza entro i livelli del predittore considerato, la differenza che si genera potrebbe riguardare solo alcune delle variabili incluse nel modello e non tutte. Al fine di verificare quali siano le variabili che variano tra i livelli del predittore, si possono considerare dei test "locali".
L'idea è quella di verificare se la distribuzione congiunta del predittore e di ognuna delle variabili qualitative sia fattorizzabile nel prodotto delle due marginali, e ci sia indipendenza o meno. Per fare ciò è necessario conoscere la distribuzione congiunta $\mathbb{P}[Y, X_k]$ ovvero $p_Y(y) \cdot \mathbb{P}[X_k|Y]$.
La distribuzione marginale della Y è simulata direttamente dal *Gibbs Sampling* per cui è disponibile. La distribuzione di una singola variabile aleatoria X_k al netto delle altre va, invece, calcolata. Considerando la distribuzione $\mathbb{P}[X_1, \ldots, X_p|Y]$, volendo ottenere la distribuzione marginale è necessario sommare su tutte le possibili combinazioni di valori che le variabili $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p$ possono assumere. Chiamando $\mathcal{A}^{-k}$ l'insieme di tutte le possibili realizzazioni del vettore aleatorio $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_P)$ si ha che

$$\begin{aligned}
\mathbb{P}[X_k|Y] &= \sum_{\mathcal{A}^{-k}} \sum_{h=1}^{H} \nu_{hy} \psi_h^{(k)} \prod_{j \neq k}^{p} \psi_{h\cdot}^{(j)} = \\
&= \sum_{h=1}^{H} \nu_{hy} \psi_h^{(k)} \underbrace{\sum_{\mathcal{A}^{-k}} \prod_{j \neq k}^{p} \psi_{h\cdot}^{(j)}}_{=1} = \\
&= \sum_{h=1}^{H} \nu_{hy} \psi_h^{(k)}
\end{aligned}$$

L'ultima uguaglianza segue dal fatto che, condizionatamente alla classe latente il le $\psi_h^{(j)}$ sono tra loro indipendenti e il loro prodotto rappresenta la distribuzione congiunta del vettore $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p$ condizionato alla classe latente, quindi sommando su tutti i pos-

sibili valori la somma è uguale ad 1.
La distribuzione marginale ricercata può essere ottenuta direttamente dall'*output* del *Gibbs Sampling*.
Per verificare se la differenza tra le due distribuzioni si può utilizzare una versione della V di Cramér basata sul modello proposta in Durante & Dunson (2014).
Per la variabile k si considera la seguente statistica

$$\begin{aligned}\rho_k^2 &= \frac{1}{\min\{M, d_k\} - 1} \sum_{y=1}^{M} \sum_{j=1}^{d_k} \frac{(\pi_{y,X_k} - p_Y \cdot \mathbb{P}[X_k = j])^2}{p_Y \cdot \mathbb{P}[X_k = j]} = \\ &= \frac{1}{\min\{M, d_k\} - 1} \sum_{y=1}^{M} \sum_{j=1}^{d_k} \frac{(p_Y \cdot \mathbb{P}[X_k = j|Y] - p_Y \cdot \mathbb{P}[X_k = j])^2}{p_Y \cdot \mathbb{P}[X_k = j]} = \\ &= \frac{1}{\min\{M, d_k\} - 1} \sum_{y=1}^{M} p_Y \sum_{j=1}^{d_k} \frac{(\mathbb{P}[X_k = j|Y] - \mathbb{P}[X_k = j])^2}{\mathbb{P}[X_k = j]}\end{aligned}$$

Sebbene la precedente statistica dia informazioni circa la differenza tra le singole variabili, un test formale del tipo:

$$H_0 : \rho_k = 0 \quad \text{vs} \quad H_1 : \rho_k \neq 0$$

non è praticabile. L'ipotesi locale puntuale può, però, essere sostituita da un'ipotesi intervallare del tipo

$$H_0 : \rho_k \leq \epsilon \quad \text{vs} \quad H_1 : \rho_k > \epsilon$$

Questa formulazione consente di calcolare agevolmente la probabilità a-posteriori dell'ipotesi nulla come la proporzione delle iterazione del *Gibbs* in cui $\rho_k \leq \epsilon$. Inoltre, come notato in Berger & Sellke (1987), questa ipotesi nulla è un'approssimazione realistica della precedente.

3.8 Importanza delle variabili

Il modello presentato consente di marginalizzare rispetto ad una qualsiasi delle variabili considerate semplicemente omettendo il vettore che la caratterizza da ogni passo del *Gibbs Samplig*. Ad esempio, il la distribuzione di probabilità condizionata senza la c-esima variabile è data da

$$\pi_y = \sum_{h=1}^{H} \nu_{hy} \prod_{j \neq c} \psi_h^{(j)}$$

Questo suggerisce la possibilità di creare una classifica delle variabili più importanti per la previsione stabilendo una qualche funzione di perdita e calcolando l'errore di previsione associato al modello senza la c-esima variabile. La variabile la cui omissione dà un errore maggiore ha un'importanza maggiore in termini di previsione. Tale tecnica definisce l'errore effettivo che si avrebbe stimando il modello omettendo dall'inizio la variabile c e non una sua approssimazione.

Capitolo 4

Simulazioni

In questo capitolo si considerano due diverse simulazioni per valutare le proprietà del modello presentato nel capitolo 3. In particolare, si pone l'accento sulla capacità del modello di effettuare i test proposti in precedenza e di riconoscere se, effettivamente, sia presente una qualche dipendenza nei dati. Si evidenzia inoltre come il modello sia in grado di stabilire l'effettivo numero di classi latenti necessarie a descrivere il fenomeno e lasciare vuote quelle in eccesso.
In uno scenario, si considera l'indipendenza delle variabili rispetto al predittore in modo da verificare se il modello identifica correttamente l'indipendenza delle variabili.
Nell'altro scenario, invece, si considera che solo alcune delle variabili disponibili presentano differenze tra i livelli del predittore, cosi da valutare anche il test proposto nella Sezione 3.7.
Per entrambe le simulazioni, si utilizza una numerosità campionaria di $n = 100$ e si considerano $p = 20$ variabili. Il limite superiore per il numero di classi latenti H è stato fissato a 25.

4.1 Scelta dei parametri per le distribuzioni a-priori

Per stimare il modello presentato nel Capitolo 3 è necessario scegliere gli iper-parametri per le distribuzioni a-priori, ovvero i parametri delle distribuzioni di Dirichlet associate alle variabili qualitative, quella associata al predittore Y (provenienza geografica) e la probabilità a-priori di generare sotto H_0, ed infine i parametri associati alla distribuzione di Dirichlet per le classi latenti.
Per il predittore Y si è adottato un approccio debolmente informativo, utilizzando come iper-parametri 0.25 per le 4 modalità.
Per il processo che determina il numero di classi latenti, si è utilizzato il reciproco del limite superiore delle classi inserite nel modello, ovvero $1/H$, nel nostro caso $1/25$ per ogni parametro, in modo da ottenere l'effetto descritto in Rousseau & Mengersen (2011).
Per l'ipotesi nulla si è preferito rimanere poco informativi e utilizzare il valore 0.5.
Infine, per le marginali delle variabili qualitative si è scelto un approccio di tipo bayesiano empirico ponendo le distribuzioni a-priori proporzionali alla frequenze osservate nel campione.

4.2 Indipendenza

Per generare i dati in modo che la distribuzione fosse indipendente dal predittore Y, ognuna delle variabili osservate è stata generata da una variabile aleatoria discreta uniforme, indipendentemente dalla classe latente. Allo stesso modo, si è generata l'appartenenza ad una delle modalità della variabile Y.
Si sono effettuate 15000 iterazioni del *Gibbs Sampler* di cui 3000 sono state considerate come *burn-in*.
Come si può notare dalla Figura 4.1 e dalla Tabella 4.1 si è sempre generato dall'ipotesi nulla, ovvero quella che considera l'indipendenza dalla Y . La probabilità a-posteriori di tale ipotesi è quindi sempre prossima ad 1.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.9999	1.0000	1.0000	1.0000	1.0000	1.0000

Tabella 4.1: Probabilità di generare dall'ipotesi di indipendenza dal predittore

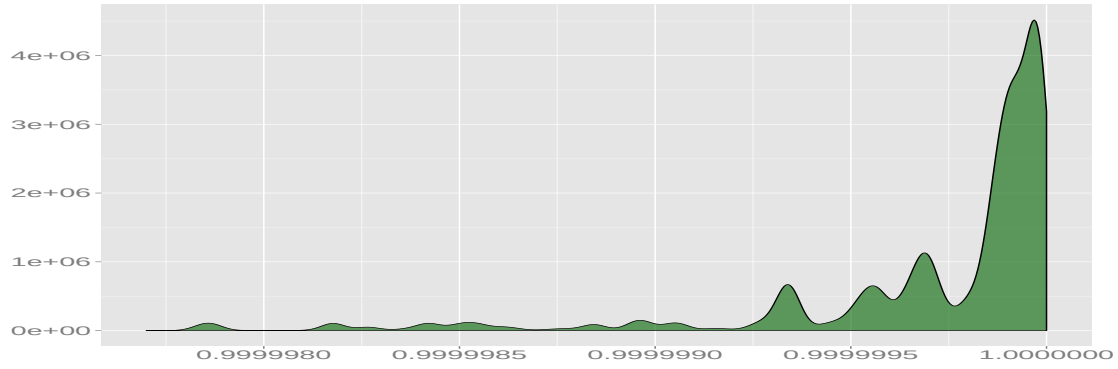


Figura 4.1: Probabilità di generare dall'ipotesi di indipendenza dal predittore

Dalla Figura 4.2 si può notare come, in ogni iterazione considerata, la percentuale di individui che appartengono alla stessa classe latente non scenda mai al di sotto del 90% circa. Tale fenomeno era ipotizzabile dal fatto che tutte le variabili condividono lo stesso meccanismo generatore.

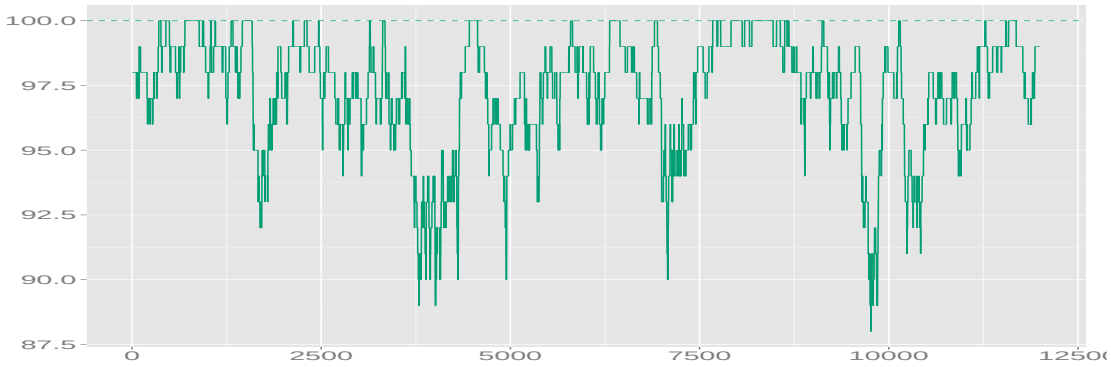


Figura 4.2: Percentuale di individui nella stessa classe in caso di indipendenza

Considerando, inoltre, di assegnare ogni individuo alla classe in cui è stato collocato più volte nel *Gibbs Sampling*, ovvero con voto di maggioranza, tutti gli individui sono stati inseriti nella stessa classe latente.

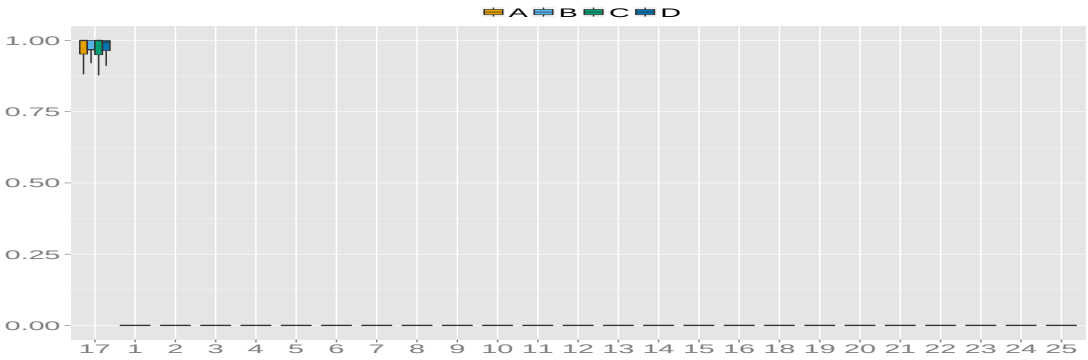


Figura 4.3: Boxplot distribuzione degli individui nelle classi in caso di indipendenza

Dal grafico (Figura 4.3) è possibile vedere come, l'a-priori utilizzata nel modello per le classi latenti ha ridotto il numero di classi ridondanti. Infatti, delle 25 classi utilizzate come limite massimo solo 1 contiene la maggior parte degli individui, mentre le altre, non necessarie a descrivere il fenomeno, rimangono vuote.

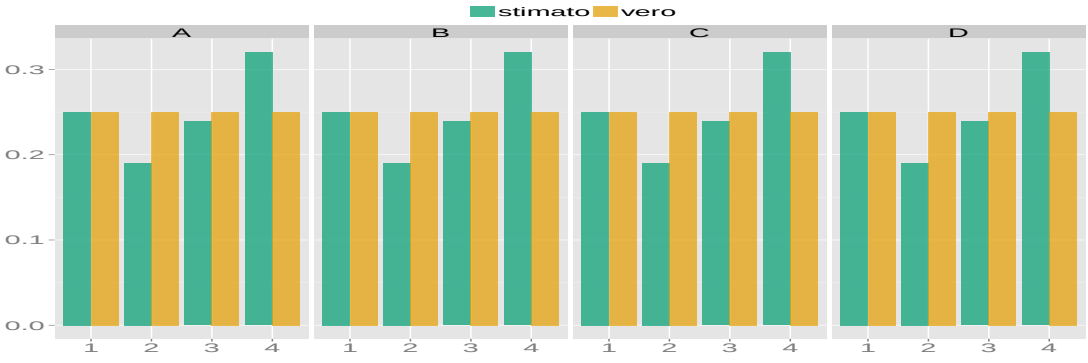


Figura 4.4: Probabilità marginali per le diverse modalità divise per i livelli del predittore in caso di indipendenza

Dalla Figura 4.4 si nota, inoltre, come utilizzando la media a-posteriori per stimare le probabilità delle variabili qualitative ordinate, al variare del predittore, sia molto prossima al vero valore dei parametri.

4.3 Dipendenza

Per simulare la dipendenza di alcune tra le variabili dal predittore Y si è considerato il seguente schema di simulazione.
Delle 20 variabili utilizzate, le prime 7 sono state generate come nel caso precedente, ovvero da una variabile aleatoria discreta uniforme, indipendentemente dal valore del predittore o dalla classe latente di appartenenza.
Per le altre variabili si è prima stabilito, per ogni individuo, il valore del predittore Y associato generandolo da una variabile aleatoria discreta uniforme. Stabilito il valore della Y si è stabilita la classe latente. Per far ciò si è posto che la probabilità di appartenere ad una delle classi latenti dipendesse dal valore della Y. In altri termini, condizionatamente al valore di Y si è stabilita la classe latente generando da una multinomiale. Si sono ipotizzate 4 classi latenti e si è posto che un individuo appartenesse alla classe latente con la stessa etichetta del predittore Y con probabilità 0.7 o in una delle restati classi con probabilità 0.1 ciascuna. Fissata per ogni individuo la classe latente, per generare i valori delle altre variabili qualitative, si è stabilito che, condizionatamente alla classe latente, le modalità della variabile venissero da una variabile aleatoria multinomiale. I parametri di queste distribuzioni sono stati generati da una Dirichlet, con parametri diversi a seconda delle classe latente di appartenenza, i parametri generati ed utilizzati per i risultati presentati sono descritti in Tabella 4.2. In particolare, delle 13 variabili considerate dipendenti, si sono considerati quattro gruppi di cui tre di tre variabili e uno di quattro. Ogni gruppo condivide gli stessi parametri della Dirichlet utilizzata per generare i parametri della multinomiale.
Come si può vedere da Figura 4.5 la distribuzione di probabilità dell'ipotesi nulla è concentratissima sullo zero, infatti, nelle 12000 replicazioni considerate non si è mai generata da essa. La più alta probabilità verificatesi è di circa 0.006.

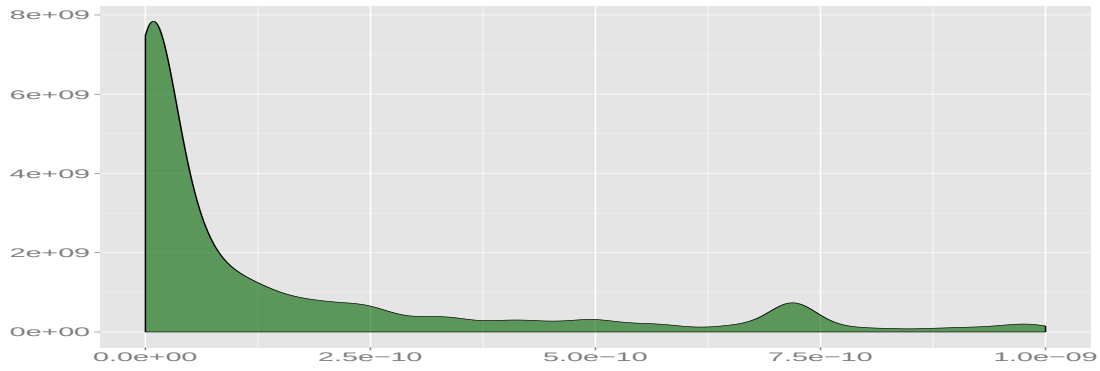


Figura 4.5: Probabilità di generare dall'ipotesi di indipendenza dal predittore

Dalla tabella Tabella 4.3 si può notare come il modello riesca perfettamente a separare

classi latenti		modalità			
		1	2	3	4
gruppo 1	1	0.814	0.053	0.063	0.070
	2	0.034	0.490	0.269	0.206
	3	0.054	0.104	0.741	0.102
	4	0.037	0.315	0.058	0.590
gruppo 2	1	0.029	0.583	0.091	0.297
	2	0.609	0.294	0.015	0.083
	3	0.099	0.136	0.041	0.724
	4	0.002	0.039	0.286	0.674
gruppo 3	1	0.176	0.133	0.111	0.581
	2	0.078	0.057	0.813	0.052
	3	0.150	0.750	0.002	0.098
	4	0.770	0.049	0.050	0.130
gruppo 4	1	0.005	0.103	0.212	0.680
	2	0.091	0.098	0.751	0.060
	3	0.062	0.614	0.204	0.120
	4	0.647	0.002	0.284	0.068

Tabella 4.2: Parametri delle multinomiali utilizzate per la simulazione

le variabili che presentano una dipendenza tra i livelli del predittore e quelle che non la presentano.

La Tabella 4.4 mostra come il modello riesca a raggruppare correttamente gli individui nelle classi latenti a cui appartengono. Infine, dalla Figura 4.6, si può notare come, anche in questo caso, le classi latenti che non sono necessarie a descrivere il fenomeno rimangano vuote.

	Classi vere			
Classi stimate	1	2	3	4
4	32	0	0	0
7	0	23	0	0
17	0	0	24	0
25	0	0	0	21

Tabella 4.4: Classi latenti vere e stimate

	val	prob
10	0.262	0.000
13	0.210	0.000
17	0.206	0.000
14	0.195	0.000
9	0.190	0.000
18	0.184	0.000
16	0.174	0.000
11	0.151	0.003
8	0.147	0.004
15	0.147	0.010
20	0.142	0.011
19	0.141	0.012
12	0.124	0.041
7	0.091	0.342
5	0.086	0.415
4	0.084	0.454
6	0.078	0.557
1	0.078	0.559
3	0.076	0.606
2	0.071	0.694

Tabella 4.3: test locale:dipendenza

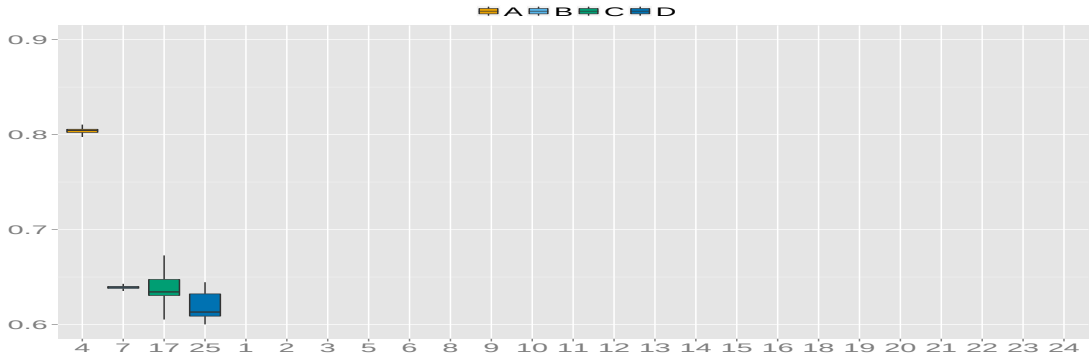
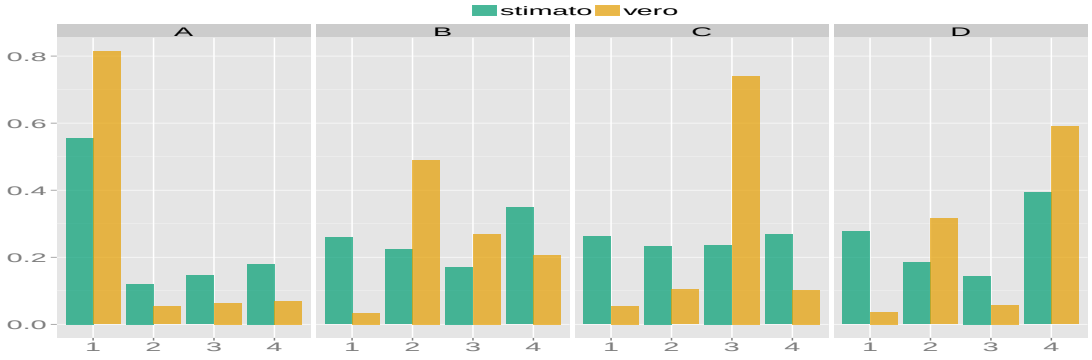
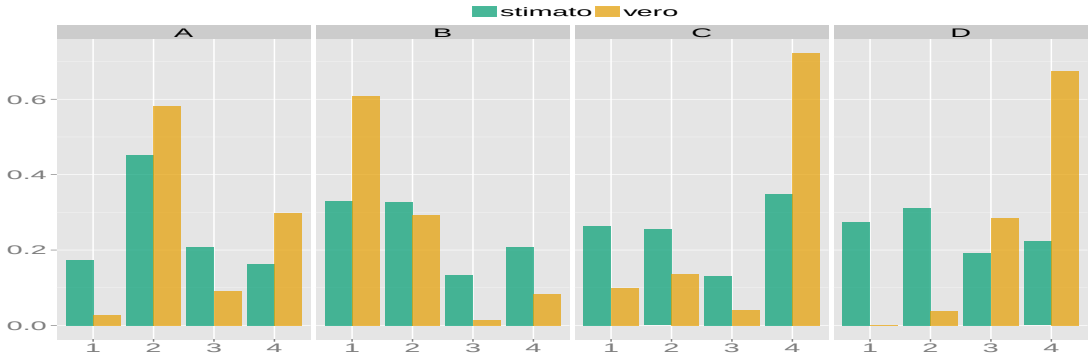


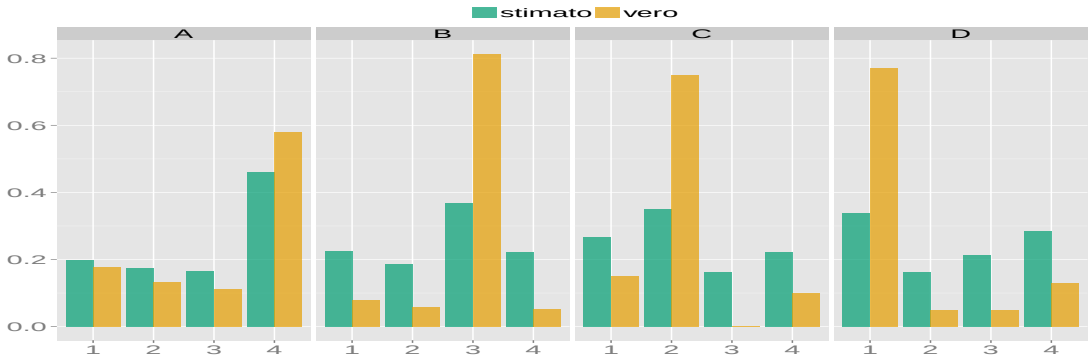
Figura 4.6: Probabilità di appartenenza alle classi latenti simulazione dipendenza



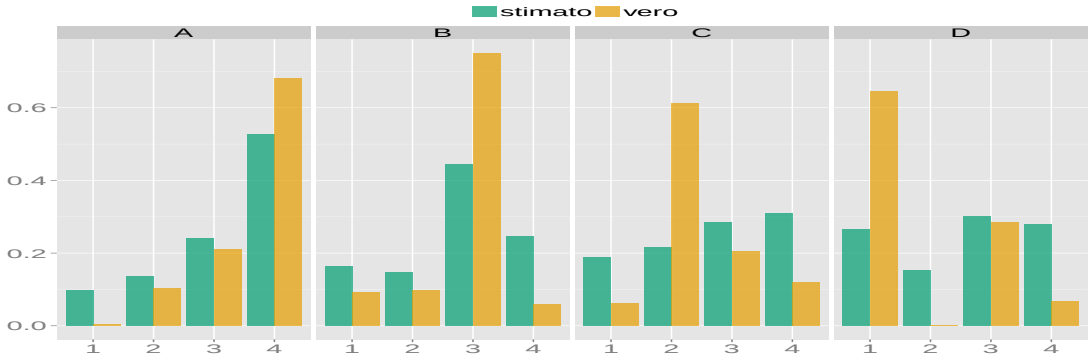
(a) Gruppo 1



(b) Gruppo 2



(c) Gruppo 3



(d) Gruppo 4

Figura 4.7: Probabilità di osservare differenti modalità nella simulazione di dipendenza

I grafici in Figura 4.7 mostrano le probabilità stimate, utilizzando la media a-posteriori, per i gruppi di variabili considerate. Sebbene la stima ottenuta non coincida perfettamente con i veri valori dei parametri esse non si discostano comunque eccessivamente.

4.4 Discussione

Dalle simulazioni, in entrambi gli scenari, si evince anzitutto che la distribuzione a-priori utilizzata per il numero di classi latenti riesce ad identificarne correttamente il numero, facendo si che le classi non necessarie rimangano vuote.
Il test globale sulle variabili, presentato nella Sezione 3.5, riesce a stabilire se globalmente esiste o meno una dipendenza dai livelli del predittore Y generando praticamente sempre dallo scenario corretto.
Analogamente il test locale, basato su una versione bayesiana della V di Cramèr, presentato nella Sezione 3.7, riesce ad identificare quali delle variabili sono differenti per i vari livelli del predittore e quali invece condividono lo stesso meccanismo generatore.
Il modello riesce a stimare, in maniera adeguata, anche la probabilità di appartenere ad una classe latente, come si può notare dai *box-plot*.
Nello scenario di indipendenza, infatti, per tutti i livelli del predittore, la probabilità di appartenere alla prima classe latente è sempre prossima ad 1. Analogamente, nello scenario in cui si è ipotizzata una dipendenza, si può notare come le mediane delle distribuzioni di probabilità per ogni classe siano vicine al vero valore 0.7.
In conclusione, le simulazioni effettuate mostrano come il modello riesca a descrivere adeguatamente i dati, sia in uno scenario in cui esiste una dipendenza, sia in uno in cui essa non è presente, discriminando correttamente le variabili che hanno un comportamento differente al variare del predittore Y.

Capitolo 5

Applicazione ai dati

In questo capitolo, si applica il modello descritto nel capitolo 3 ai dati disponibili sugli odori. Come detto nel capitolo 1, il modello è stato stimato su 246 dei 328 soggetti disponibili, pari al 75% della numerosità totale in modo da poter valutare la bontà di previsione del modello. Per effettuare l'inferenza a-posteriori, si sono effettuate 15000 estrazioni dal *Gibbs Sampler*, proposto nel capitolo 3, di cui le prime 3000 sono state scartate come *burn-in*.
Per ottenere le stime ci sono voluti circa 40 minuti utilizzando una *workstation* portatile con processore *Intel® Core™ i7-4710MQ CPU @ 2.50GHz* con 16Gb di RAM. Il codice quasi totalmente in linguaggio R, con qualche funzione scritta in *C++*.
Per la scelta degli iper-parametri delle distribuzioni a-priori si è utilizzata la stessa logica del Capitolo 4.
In particolare, per la scelta della probabilità a-priori dell'ipotesi di indipendenza sul giudizio degli odori dalla provenienza geografica si è preferito rimanere poco informativi, nonostante la presenza di studi su tale effetto in altre aree del mondo.

5.1 Convergenza del modello

Come si può notare dal grafico 5.1, i soggetti appartenenti alla stessa area geografica, vengono inseriti nella stessa classe latente e non c'è sovrapposizione tra i vari gruppi. Questo indica che la dipendenza dall'area geografica è effettivamente espressa dall'appartenenza ad una specifica classe latente. Si può notare come molte delle classi rimangano vuote, segno che il numero massimo di classi H posto uguale a 25 è sufficiente ad approssimare il processo di Dirichlet.
Per valutare se il modello fosse effettivamente giunto a convergenza si sono ispezionati i *trace-plot* relativi alla probabilità di un individuo di appartenere ad una area geografica e quelli relativi alle probabilità di osservare una modalità sempre divisa per zona geografica. Queste quantità sono quelle utilizzate per l'inferenza successiva. Tutti i *trace-plot* presentano un buon *mixing*, segno che le distribuzioni marginali sono giunte a convergenza.
Sono riportati, a titolo di esempio, i grafici relativi alla probabilità di un individuo tra i 246 utilizzati per la stima di appartenere ad una data zona geografica (Figura 5.2) e quelli relativi alla probabilità di osservare una data modalità per il *rum* (Figura 5.3). I trace plot sono

stati disegnati utilizzando le ultime 12000 unità del campione e la linea gialla rappresenta, in entrambi i casi, la media cumulata fino all'iterazione.

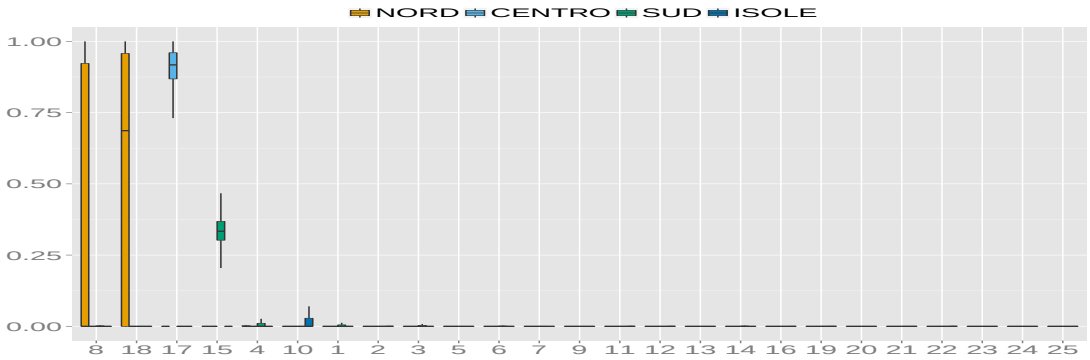


Figura 5.1: boxplot: probabilità di appartenere ad una classe latente distinta per provenienza geografica

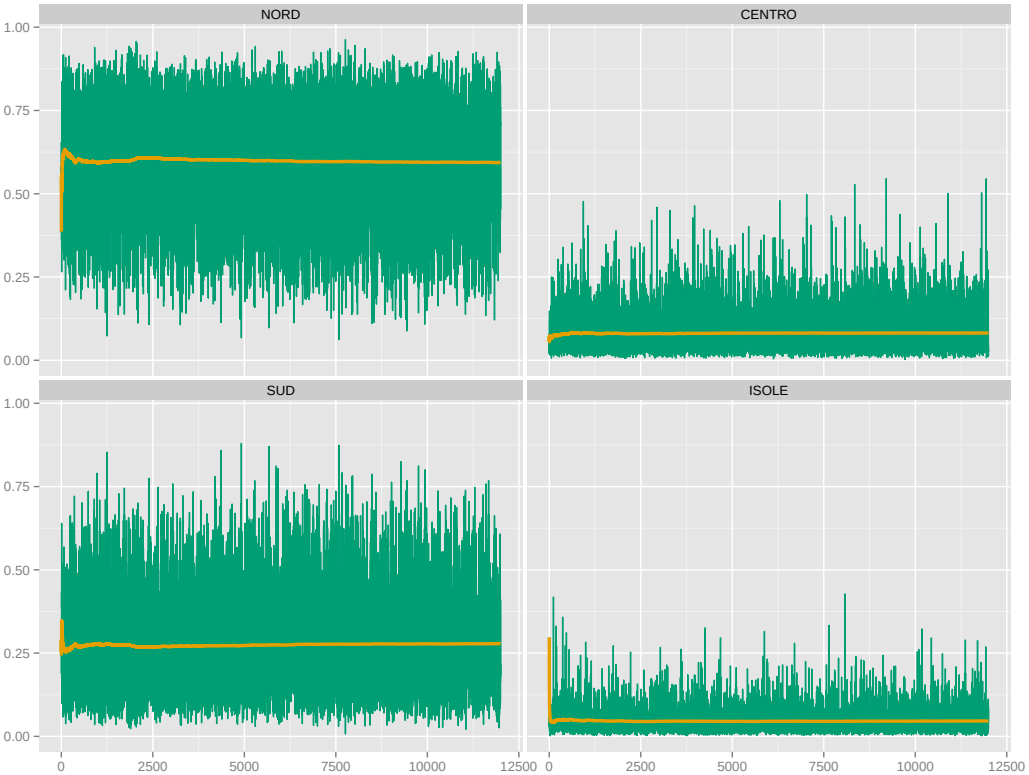


Figura 5.2: trace-plot: probabilità di appartenere ad una delle 4 zone considerate

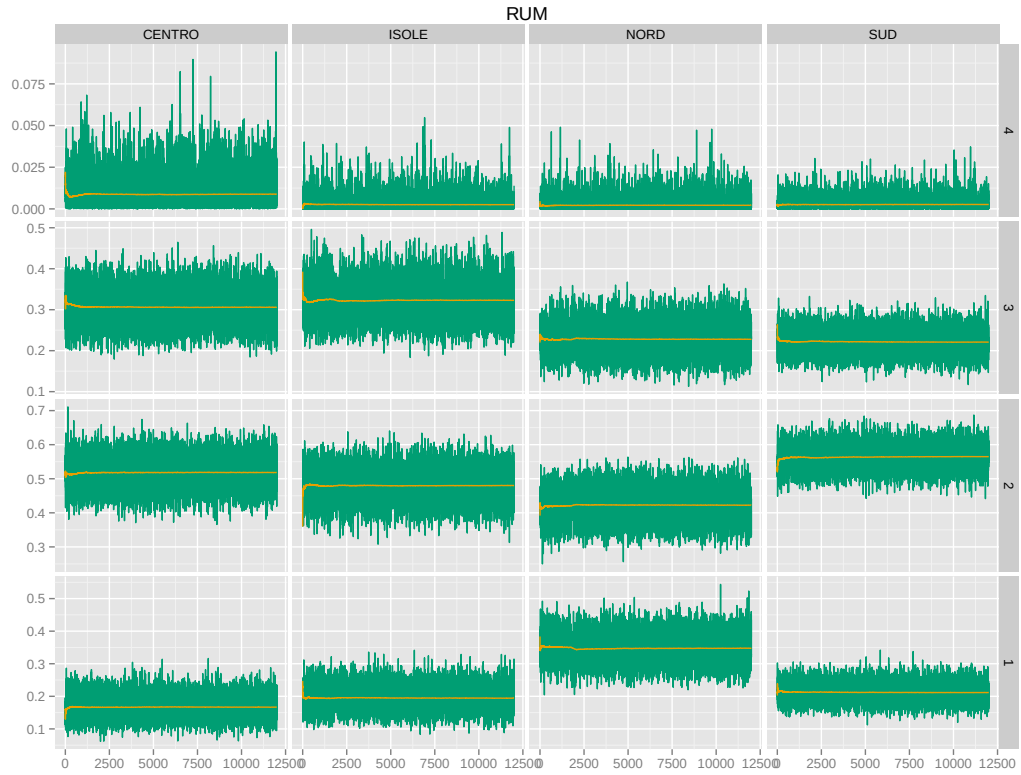


Figura 5.3: trace-plot: probabilità di osservare differenti modalità divisa per zone geografiche per il *rum*

5.2 Test

Il modello considerato include la mistura di due possibili ipotesi, dove l'ipotesi H_0 indica che non esiste una differenza significativa tra la percezione degli odori nelle differenti regioni italiane.
Nelle 12000 iterazioni considerate, dopo il *burn-in*, non si è mai generato dall'ipotesi H_0 segno di una forte evidenza empirica a favore dell'ipotesi alternativa. La probabilità a-posteriori di H_0, per ogni iterazione del *Gibbs Sampler*, è sempre prossima allo 0 (il valore più grande ottenuto è di circa $2.83 \cdot 10^{-93}$).
Nella tabella 5.1 sono, invece, presentati i risultati del test, basato sulla V di Cramèr, presentato nella Sezione 3.7. Per la probabilità a-posteriori di H_0, si è utilizzata l'approssimazione proposta alla fine della Sezione 3.7 ponendo $\epsilon = 0.05$.
Si può notare come l'ordinamento degli odori, in base alla probabilità a-posteriori di H_0, non differisca eccessivamente da quello ottenuto applicando il test χ^2 di Pearson presentato in tabella 1.3. Qualche odore occupa una posizione differente in classifica: ad esempio si veda *sour cherry* che nel test χ^2 risulta in posizione più elevata e *oleic acid* che invece si trova in una posizione meno elevata nella classifica. Si può inoltre notare come molte delle variabili condividano la stessa media a-posteriori e/o lo stessa probabilità che l'ipotesi nulla sia falsa, formando dei gruppi di variabili che hanno un comportamento simile.
Oltre a stabilire quali variabili presentano delle differenze tra le varie regioni può essere di interesse stabilire tra quali regioni e tra quali modalità sono presenti queste differenze. Dal *Gibbs Sampling* sono disponibili le distribuzioni a-posteriori del verificarsi di tutte le modalità per ogni odore e per ogni area geografica.
Un modo per effettuare tale test è considerare le estrazioni dal *Gibbs Sampling* come un campione indipendente da una data variabile aleatoria e confrontare per ogni modalità le distribuzioni del nord, centro, sud, isole.
Per effettuare il test si è utilizzato il test di Kolmogorov-Smirnov a due campioni. La statistica del test quantifica la distanza tra due funzioni di distribuzioni empiriche. L'ipotesi nulla del test è che i due campioni sottoposti a test provengano dalla stessa popolazione. Questo test è uno dei test non parametrici più generali per confrontare due campioni dato che riesce a considerare differenze sia nella forma che nella scala delle distribuzioni.
La statistica test è data dall'estremo superiore del valore assoluto delle differenze delle due funzioni di ripartizione empiriche e la distribuzione, sotto l'ipotesi nulla, non dipende dalla vera funzione di ripartizione delle variabili sottoposte a test. Nella Tabella 5.2 sono riportate le statistiche calcolate e i valori-p per i confronti a coppie del *rum*. Si può notare che non in tutte le modalità si evidenziano delle differenze. Ad esempio, nel caso del *rum* non c'è differenza tra gli idividui che considerano il rum molto spiacevole, come si può evincere anche dal grafico in Figura 5.4.

	media a-posteriori	probabilità
propionic acid	0.15	0.00
caproic acid	0.14	0.00
almond	0.13	0.00
valeraldehyde	0.11	0.00
rum	0.13	0.00
acetic acid	0.10	0.00
propylene glycol	0.10	0.00
fish composition	0.10	0.01
valeric acid	0.10	0.01
ethyl propionate	0.10	0.01
benzaldehyde	0.10	0.01
oleic acid	0.10	0.01
siberian musk deer	0.10	0.01
diesel fuel	0.10	0.01
n butanol	0.10	0.01
octanoic acid	0.09	0.01
strawberry	0.10	0.01
pineapple	0.11	0.01
cyklohexanone	0.09	0.01
deer	0.09	0.01
ethyl acetate	0.09	0.01
water	0.09	0.01
formic acid	0.09	0.02
n butanol2	0.08	0.02
fishing cat	0.09	0.02
men's parfume	0.08	0.03
buru babirusa	0.08	0.03
women's parfume	0.08	0.04
coconut	0.08	0.05
lemon	0.08	0.06
sour cherry	0.08	0.07
vanilla	0.07	0.14

Tabella 5.1: Risultati test locale

	1	2	3	4
	0.115	0.936	0.370	0.000
NORD-CENTRO	**0.015**	**0.007**	**0.012**	0.357
	0.799	0.482	0.123	0.000
NORD-SUD	0**.008**	**0.011**	**0.015**	0.327
	0.694	0.799	0.976	0.000
NORD-ISOLE	**0.009**	**0.008**	**0.006**	0.199
	0.218	0.291	0.855	0.000
CENTRO-SUD	**0.014**	**0.013**	**0.008**	0.164
	0.502	0.575	0.321	0.000
CENTRO-ISOLE	**0.011**	**0.010**	**0.012**	0.326
	0.952	0.586	0.056	0.000
SUD-ISOLE	**0.007**	**0.010**	**0.017**	0.282

Tabella 5.2: Testi di Kologorow-Smirnov per *rum*

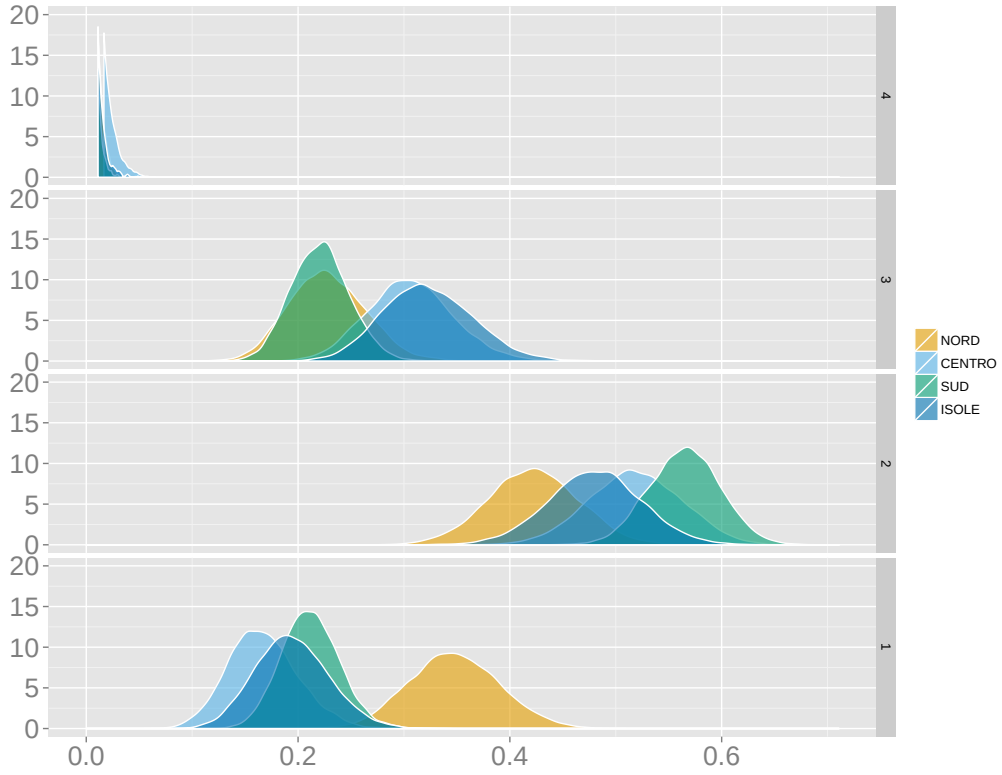


Figura 5.4: Densità a-posteriori dei giudizi sul *rum*

5.3 Confronto con altri modelli di previsione

In questa sezione si è confrontata l'inferenza ottenuta con il modello proposto, in particolare la capacità previsiva con alcuni tra i modelli utilizzati in letteratura per problemi analoghi. Per ottenere la previsione dal modello del Capitolo 3 si sono utilizzati due differenti metodi: il primo consiste nel ottenere, per ogni iterazione del *Gibbs Sampler* la probabilità che l'individuo i appartenga ad una delle classi proposte e, considerando come previsione la moda a-posteriori si è proceduto ad assegnare una classe con voto di maggioranza. Una seconda metodologia è stata quella di considerare, ottenuta la distribuzione di appartenenza ad ogni classe, il massimo a-posteriori per ogni modalità ed assegnare, ad ogni individuo la classe con massimo osservato maggiore.

5.3.1 Valutazione dei modelli

Da un punto di vista statistico la valutazione di un modello solo sulla base della previsione è un criterio discutibile. Quando ci si trova in un contesto sperimentale bisogna tener conto delle condizioni dell'esperimento e far si che le assunzioni del modello siano coerenti con l'esperimento stesso.
In ambito pratico, quando non si ha controllo sull'esperimento, raramente si possono applicare considerazioni di tipo esclusivamente teorico, in quanto ogni modello è sbagliato e quindi include sia una componente di varianza che una di distorsione. Queste due quantità sono in contrasto è quindi necessario trovare un compromesso. In contesti di questo tipo, soprattutto quando si vuole prevedere una variabile di interesse, è prassi considerare l'errore associato al modello su un insieme di dati non utilizzato nella fase di stima. Sebbene non posso fornire valutazioni assolute sul modello, questa tecnica è utile per confrontare modelli diversi e dare indicazioni sulla scelta del modello più opportuno ad un dato problema. Questo metodo deriva dal principio del campionamento ripetuto, si immagina, infatti, di disporre di due campioni generati dallo stesso meccanismo. Il fatto di non utilizzare lo stesso insieme di dati per la parte di stima del modello e per il confronto dello stesso con altri deriva dal fatto che l'insieme utilizzato per la stima dà una previsione ottimistica di quello che è il risultato, premiando modelli che si avvicinano molto ai dati e che non considerano l'errore di distorsine del modello.
In ambito bayesiano,si fa riferimento ad un'impostazione soggettiva, per cui , la distribuzione a-posteriori è condizionata allo specifico campione non è quindi corretto, da un punto di vista "filosofico", valutarlo su un campione differente da quello utilizzato per la stima del modello stesso.
Se l'obiettivo è quello di confrontare diversi modelli il metodo "corretto" sarebbe quello di calcolare la probabilità a-posteriori che un dato modello abbia generato il campione in esame. Da un punto di vista pratico, però, l'informazione sulla capacità previsiva dà comunque indicazioni su quale modello possa essere preferibile. In sintesi, per confrontare i modelli e le previsioni, ottenute con procedure bayesiane e frequentiste tra loro ci affidiamo al principio del campionamento ripetuto (facendo quindi una forzatura nei confronti dell'impostazione

bayesiana ortodossa) e in quest'ottica confrontiamo i risultati di previsione ottenuti su un insieme diverso da quello usato per la stima dei modelli.

5.3.2 Modello Multinomiale

Il modello più simile a quello proposto, largamente usato in letteratura per l'analisi delle variabili qualitative quando una è di interesse come risposta, è il modello multinomiale. Tale modello ha una stretta connessione con i modelli log-lineari; infatti, un risultato noto è che la distribuzione condizionata di un inseme di variabili aleatorie di Poisson, fissato il totale ha una distribuzione multinomiale con vettore delle probabilità dato dal rapporto tra le medie delle Poisson e il totale.
L'interesse in questo modello è quello di stimare la probabilità di appartenere ad una data modalità della variabile risposta condizionatamente ai valori assunti dalle altre variabili osservate.
Più formalmente, data una variabile risposta Y con M livelli $\{1 \dots, M\}$ ed un insieme di variabili esplicative z, la probabilità di appartenere ad uno dei gruppi può essere espressa come

$$\mathbb{P}(Y = k|Z = z) = \frac{e^{\beta_{0k}+\beta_k^T z}}{\sum_{\ell=1}^{M} e^{\beta_{0\ell}+\beta_\ell^T z}}$$

definita A la matrice $n \times p$ il cui generico elemento $a_{i\ell} = \mathbb{1}(y_i = \ell)$ la funzione di verosimiglianza del modello può essere espressa come

$$\ell(\{\beta_{0k}, \beta_k\}_1^M) = -\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{l=1}^{M} a_{il}(\beta_{0k} + x_i^T\beta_k) - \log\big(\sum_{l=1}^{M} e^{\beta_{0k}+x_i^T\beta_k}\big)\Big)\right]$$

Dove, β è una matrice di coefficienti $p \times M$. Il termine β_k si riferisce alla k-esima colonna (categoria k della variabile risposta) e β_j alla j-esima riga (vettore di M coefficienti per la variabile j). Per la stima del modello si è utilizzata la funzione *multinom* del pacchetto *nnet* di *R*, presentata in Venables & Ripley (2013).
Data la presenza di molte variabili si è inoltre applicata una procedura *step-wise* in modo da selezionarne solo alcune. In particolare, si è utilizzata una procedura *backward* basata sull'*AIC*, ovvero partendo dal modello con tutte le variabili si sono eliminate in sequenza le variabili che portassero ad *AIC* minore e fermandosi al modello in cui l'*AIC* non potesse essere migliorato togliendo una delle variabili ancora presenti nel modello.

5.3.3 Lasso Multinomiale

Alternativamente alla selezione delle variabili presentata nella sezione precedente si può introdurre una penalizzazione dei coefficienti. In particolare, si è utilizzata una penalizzazione di tipo *lasso* (Friedman et al., 2010).
Seguendo la notazione del paragrafo precedente, la funzione di verosimiglianza penalizzata

del modello può essere espressa come

$$\ell(\{\beta_{0k}, \beta_k\}_1^M) = -\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{l=1}^{M} a_{il}(\beta_{0k} + x_i^T\beta_k) - \log\big(\sum_{l=1}^{M} e^{\beta_{0k}+x_i^T\beta_k}\big)\Big)\right] + \lambda\sum_{j=1}^{p}||\beta_j||_q$$

Dove, il termine di penalizzazione $||\beta_j||_q$ presenta due possibili opzioni per $q \in \{1, 2\}$. Quando $q = 1$, si ottiene una penalizzazione lasso per tutti i parametri, ovvero ogni parametro del modello può essere nullo o meno. Quando $q = 2$, si ha una penalità così detta *grouped-lasso*, ovvero tutti gli M coefficienti di una particolare variabile possono essere o tutti nulli o meno. Per la stima del modello si è utilizzato l'algoritmo *partial Newton algorithm*, implementato nel pacchetto *glmnet* di *R* (Friedman et al., 2009). L'algoritmo è basato su un'approssimazione quadratica della log-verosimiglianza che consente al vettore di parametri (β_{0k}, β_k) di variare per una sola classe alla volta. Per ogni valore di λ si fa un ciclo dell'algoritmo calcolando l'approssimazione quadratica della log-verosimiglianza per ogni classe.
Per selezionare il parametro di penalizzazione λ, sia per il modello *grouped* che per il modello *ungrouped*, si è utilizzata la convalida incrociata. L'insieme di stima è stato diviso in 10 parti di circa uguale numerosità e si è stimato il modello su 9 di esse utilizzando la restante parte per la previsione, cambiando di volta in volta la porzione di dati su cui effettuare la previsione. Per ogni ciclo si è, quindi, calcolato l'errore di classificazione totale, ottenendo alla fine una stima della distribuzione di tale errore di cui si è calcolata la media e la deviazione standard (Figura 5.5). Le due linee tratteggiate indicano il λ che minimizza l'errore di classificazione e il più grande λ tale che l'errore sia compreso nell'intervallo dato da una volta la deviazione standard dal minimo (usato per la stima finale).

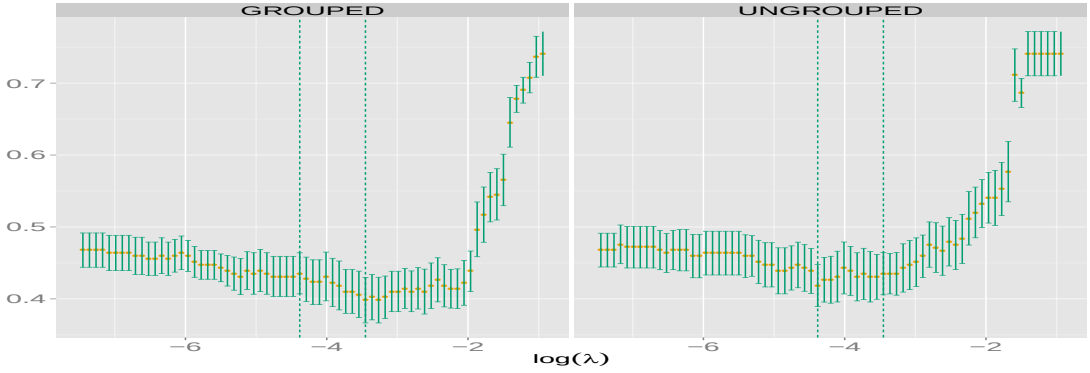


Figura 5.5: Errore di classificazione convalida incrociata: parametro lambda il lasso multinomiale

5.3.4 Foreste Casuali

Per le *Foreste Casuali* si è fatto riferimento a quanto proposto in Breiman (2001). Le *Foreste Casuali* generalizzano il *Bagging*, una tecnica per la riduzione della varianza nella previsione. In generale, per modelli con varianza elevata e distorsione piccola il *Bagging* funziona particolarmente bene: questo lo rende indicato per modelli come gli alberi.
Nell'ambito della classificazione, la procedura consiste nell'estrarre un campione *bootstrap* e stimare un albero per ogni campione. Stimati gli alberi per ognuno di essi, viene effettuata la

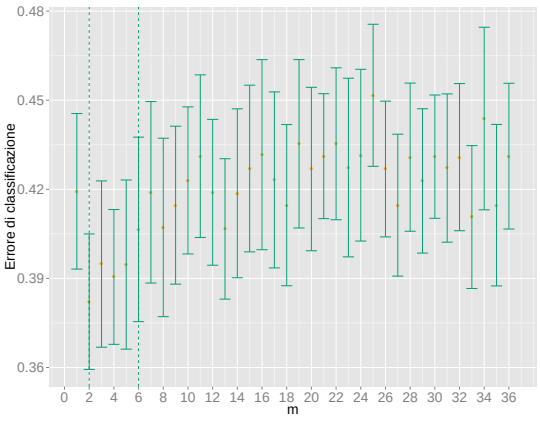
previsione e quella totale viene stabilità con voto di maggioranza. Sebbene stimati da campioni indipendenti, gli alberi provenienti da tale procedura sono tra di loro molto correlati. Le *Foreste Casuali* tentano di ridurre la correlazione tra gli alberi senza aumentarne eccessivamente la varianza, in modo da ridurre l'errore complessivo. L'algoritmo per la stima del modello prevede i seguenti passi

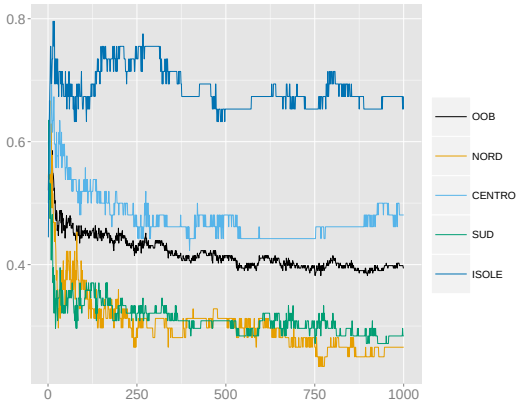
1. Per $b = 1$ a B, dove B è il numero di campioni *bootstrap*:
   (a) Si estrae un campione *bootstrap* Z^* di numerosità n dall'inseme di stima;
   (b) Si crea una *Foreste Casuali* T_b dal campione *bootstrap* ripetendo ricorsivamente i seguenti passaggi per ogni nodo terminale sull'albero finché non si raggiunge una taglia minima prefissata.
      i. Si selezionano m variabili tra le p variabili disponibili.
      ii. Si selezionano la migliore variabile per ogni *split* tra le m selezionate
      iii. Si divide il nodo in due nodi discendenti.
2. Si fornisce come risultato l'insieme di alberi $\{T_b\}_1^B$

   Per classificare un nuovo punto x sia $\hat{C}_b(x)$ la classe prevista dalla b-esima foresta casuale allora la previsione totale è data da $\hat{C}_{rf}^B(x) =$ voto di maggioranza $\{\hat{C}_b(x)\}_1^B$

Una caratteristica importante delle *Foreste Casuali* è l'uso dell'errore *out-of-bag* Per ogni osservazione si costruisce la previsione solo su quegli alberi stimati sui campioni *bootstrap* per i quali l'osservazione non è presente. La stima dell'errore *out-of-bag* è, in generale, quasi identica a quella ottenuta dalla convalida incrociata, quindi, al contrario di altri metodi di stima non lineari, le *Foreste Casuali* possono essere stimati in una unica sequenza con una sorta di convalida incrociata effettuata durante il processo di stima. Una volta che l'errore *out-of-bag* si è stabilizzato la fase di stima del modello può terminare.
Un'altra caratteristica interessante delle *Foreste Casuali* è quella di poter fornire un' idea di quali siano le variabili più importanti tra quelle disponibili. Come per ogni modello basato sugli alberi, il miglioramento nel criterio utilizzato per la costruzione è una misura d'importanza attribuita alla variabile, e può essere cumulata per ogni albero della foresta casuale separatamente per ogni variabile. Esiste, inoltre, un metodo alternativo per il calcolo di tale importanza valido per le *Foreste Casuali*. Quando si è nella fase di crescita del b-esimo albero si effettua la previsione per il campione *out-of-bag* dopo di che i valori per la j-esima variabile sono permutati in maniera casuale e si ricalcola la previsione. La perdita nell'accuratezza della previsione è usata come misura di importanza della variabile j nella *Foreste Casuali* Contrariamente a quanto proposto nella Sezione 3.8, questo metodo non rappresenta l'effetto sulla previsione che si avrebbe se la variabile non fosse usata in fase di previsione, questo perché se il modello fosse ristimato senza quella variabile, un'altra potrebbe essere usata come surrogato. L'importanza associata alle variabili è presentata nella Figura 5.6
Per fissare il parametro m delle *Foreste Casuali* si sono seguite due differenti metodologie, si è tarato il parametro utilizzando una convalida incrociata con 10 gruppi come per il parametro

di penalizzazione del lasso e si è poi seguita la raccomandazione dell'autore, ovvero si è posto $m = \lfloor \sqrt{p} \rfloor$





	1	2	3	4	5
modello multinomiale	0.0000	0.0000	0.0000	0.0000	0.0000
modello multinomiale (stepwise)	0.0000	0.0000	0.0000	0.0000	0.0000
modello proposto (voto di maggioranza)	0.2114	0.2361	0.1961	0.2000	0.2105
modello proposto (massimo a-posteriori)	0.1992	0.2143	0.1837	0.2048	0.1818
Lasso Multinomiale (ungrouped)	0.0813	0.0606	0.0755	0.0864	0.1087
Lasso Multinomiale (grouped)	0.1463	0.1077	0.1132	0.1860	0.1667
Foreste Casuali (m = 6)	0.3943	0.2985	0.3721	0.4257	0.5143
Foreste Casuali (m = 2)	0.3943	0.3425	0.2857	0.4545	0.4286

Tabella 5.3: tabella errori: insieme di stima

	Totale	Nord	Centro	Sud	Isole
modello multinomiale	0.5732	0.8571	0.5385	0.3913	0.6316
modello multinomiale (stepwise)	0.5732	0.8571	0.5385	0.3913	0.6316
modello proposto (voto di maggioranza)	0.4634	0.6667	0.4839	0.2500	0.4783
modello proposto (massimo a-posteriori)	0.4268	0.5833	0.4643	0.2800	0.4706
Lasso Multinomiale (ungrouped)	0.4268	0.6471	0.4000	0.2500	0.5000
Lasso Multinomiale (grouped)	0.3659	0.6000	0.3600	0.2903	0.2727
Foreste Casuali (m = 6)	0.3293	0.5714	0.3333	0.1852	0.3571
Foreste Casuali (m = 2)	0.3293	0.6364	0.2727	0.2857	0.2857

Tabella 5.4: tabella errori: insieme di verifica

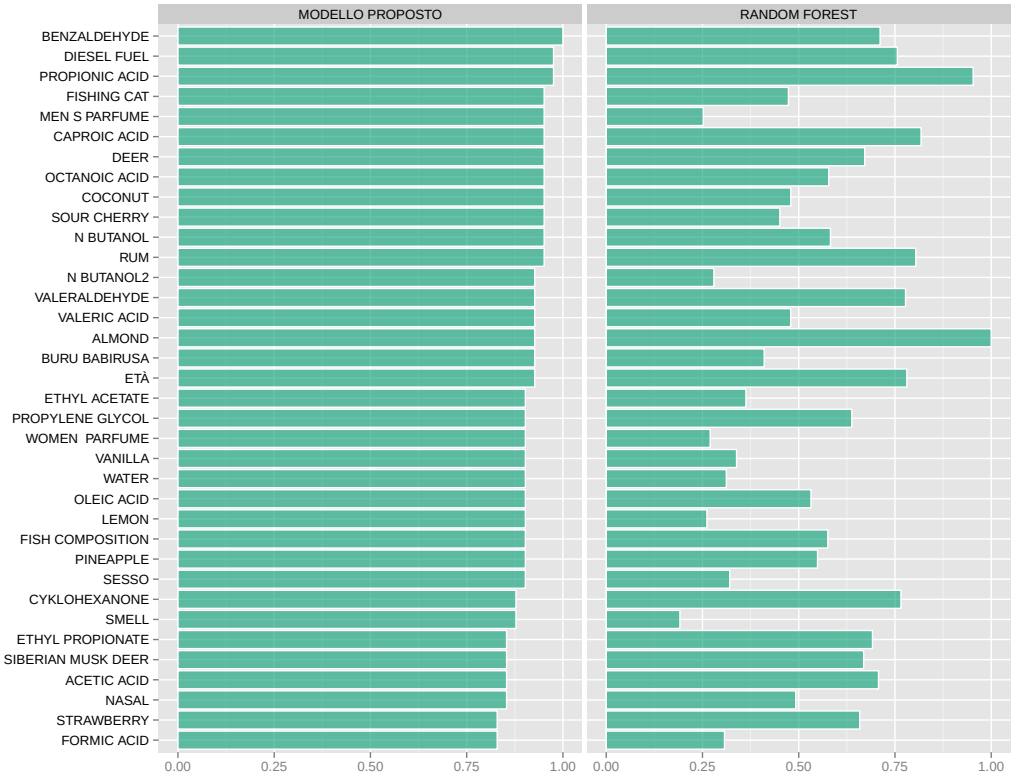


Figura 5.6: Importanza delle variabili

5.4 Discussione

La valutazione, in termini di previsione, è stata effettuata su solo 82 soggetti: non ha, per cui, un accuratezza elevatissima.
Come si può notare dalla Tabella 5.3 il modello multinomiale, sia con selezione *step-wise* che senza, presenta un errore nullo nell'insieme di stima. Tale condizione è sintomo di sovradattamento (*overfitting*) ovvero il modello, presentando troppi parametri, ristima totalmente i dati e non ha sufficiente capacità di astrazione per spiegare qualcosa sul fenomeno in esame. Questo risultato è confermato dal fatto che nell'insime di verifica (Tabella 5.4) i modelli multinomiali sono quelli con un errore di previsione peggiore. I due modelli presentano di fatto lo stesso errore sull'insieme di verifica, segno che la procedura *forward* adottata non è sufficiente a selezionare in maniera adeguata le variabili.
I due modelli lasso funzionano, al contrario, molto meglio in termini di previsione. In particolare, il modello multinomiale con penalizzazione lasso *grouped* ha un errore non molto maggiore delle *Foreste Casuali* che sono i modelli, tra quelli proposti, che prevedono meglio. Il modello ha, inoltre, l'errore più piccolo tra quelli riscontrati per quel che riguarda la corretta previsione delle isole.
Le *Random Forset*, anche utilizzando il parametro di *default*, ovvero, senza alcuna taratura, presentano l'errore di classificazione più basso in assoluto. Questi modelli sono, però, basati su una combinazione di classificatori ed offrono, quindi, minori capacità in termini di interpretazione dei modelli multinomiali. In particolare, il modello lasso *grouped* non sembra avere un errore molto maggiore in termini assoluti delle *Foreste Casuali* potrebbe, quindi, comunque essere preferibile.
Il modello proposto presenta errori di classificazione leggermente maggiori di quelli del modello lasso *grouped* ma comunque simili a quello del modello lasso *ungrouped*. Il modello non inserisce, comunque, nessuno *step* di selezione, mediante il quale probabilmente si potrebbero migliorare le capacità previsive.
Per quel che riguarda l'importanza delle variabili al fine di poter effettuare un confronto, le misure sono state normalizzate. Per il modello proposto l'importanza è stata calcolata sull'insieme di verifica, mentre, per le *Foreste Casuali* si è sfruttato l'errore *out-of-bag*. Dal grafico (Figura 5.6) si può notare come la *Foreste Casuali* e il modello proposto diano risultati discordanti.
Le prime 6 variabili, considerate più importanti dalla *Foreste Casuali*, sono le stesse considerate maggiormente significative dal test di Pearson (Tabella 1.3). Altre variabili occupano posizioni differenti ma i risultati sono comunque coerenti con quelli ottenuti nei test effettuati nel Capitolo 1. Nel modello proposto, la classificazione è coerente con quella del test basato sulla V di Cramèr, infatti, anche in questo caso si evincono blocchi di variabili a cui viene associata la stessa importanza.

Conclusioni

In questa tesi si è presentato un modello bayesiano non parametrico per lo studio della distribuzione di probabilità congiunta di una tabella di contingenza quando quest'ultima presenta una dipendenza da una variabile esplicativa qualitativa.
Per quel che riguarda l'applicazione ai dati si può concludere che, come era prevedibile dagli studi condotti in altre parti del mondo, esistono delle differenze nel giudizio attribuito a differenti odori. Inoltre, tali differenze riguardano la maggior parte degli odori sottoposti a test, segno che con buona probabilità fattori ambientali e culturali influiscono notevolmente sulle percezioni olfattive.
Il modello presenta, in fase di simulazione, delle buone caratteristiche riuscendo a stabilire se la variabile esplicativa considerata influenzi o meno la misura di probabilità e a collocare perfettamente gli individui nelle classi di appartenenza.
L'applicazione ai dati ha fornito dei risultati soddisfacenti anche in termini di previsione. Pur non essendo il modello che prevede meglio, quello proposto raggiunge risultati simili a modelli che effettuano selezione delle variabili. Questo suggerisce che, il modello potrebbe essere migliorato inserendo un passo di selezione. Alternativamente, si potrebbe considerare che alcune delle variabili siano indipendenti non solo condizionatamente alle classi latenti ma anche in generale portando così ad una rappresentazione della tabella di contingenza con una dimensione inferiore ed auspicabilmente a prestazioni migliori in termini di precisione.
Per i test sarebbe inoltre possibile estendere la metodologia proposta in modo che non tenga conto solo delle distribuzioni marginali ma anche dell'effetto di tutte le altre variabili aleatorie osservate.

Appendice A:
Codice sviluppato

Gibbs Sampling

```
nrep
PSI = list()
NUH = list()
ZI =  list()
PT = list()
TT = list()
PY = list()
PROB = list()

set.seed()
cat("for is starting.....","\n")

ptm = proc.time()
for(b in 1:nrep)
{
  # 1) aggiorniamo le Psi
  datih = lapply(1:H,function(x) subset(dati,zi==x))
  count_psi = rapply(datih, function(x) table(x),how="list")
  # parametri della psi
  par_psi = lapply(1:H, function(h)
                          mapply(function(x, y)
                                 x+y, count_psi[[h]], prior_psi))

  psi = PSI[[b]]= rapply(par_psi,function(x)
                                  gtools::rdirichlet(1,x),how="list")


  # 2) Generiamo le classi latenti per ogni individuo
  probs = PROB[[b]] =get_prob_c(psi,nuh,yy,dati2)
  zi = ZI[[b]]=
  sapply(1:NROW(dati),function(x)
                        resamp(1:H,replace=TRUE,size=1,prob=probs[x,]))
```

```
  # 3) generiamo T dalla bernoulliana corrispondente
  n = sapply(1:H, function(x) sum(zi==x))
  # ogni riga è per una modalità di y
  ny =t(sapply(1:length(levels(y)),function(m)
                                        sapply(1:H, function(x)
                                                sum(zi[y==m]==x))))

  pT =PT[[b]] = test_prob(ph0,prior_u,n,ny)
  T =TT[[b]] = rbinom(1,1,1- pT)

  # 4) generiamo i pesi nel caso in cui si sta generando da H0

  if(T==0)
  {
    nuh =  as.vector(gtools::rdirichlet(1,prior_u + n))
    # lo ripeto 4 volte per mantenere la stessa struttura
    nuh = NUH[[b]] = t(sapply(1:length(levels(y)),function(m) nuh))
  }

  #4) generiamo i pesi nel caso in cui si sta generando da H1

  else
  {
    # ogni riga è per una modalità di y
    nuh = NUH[[b]] =
    t(apply(ny,1,function(m)
                  as.vector(gtools::rdirichlet(1,prior_u + m ))))

  }

  # generiamo la y
  PY[[b]] = as.vector(gtools::rdirichlet(1,post_y))

  if(b %% 100 ==0) cat("Iteration = ",b,"\n")

}
tot_time =  proc.time() -ptm
tot_time
cat("for ended....","\n")
```

Funzioni R richiamate nel Gibbs

```
# se la lunghezza del vettore in input è 1 ritorna il numero stesso

resamp = function(x,...){if(length(x)==1) x else sample(x,...)}


# logaritmo della funzione beta multivariata

lmbeta = function(x)
{
  S = sum(x)

  out = sum(lgamma(x)) - lgamma(S)
  return(out)
}

# funzioni per la distribuzione a-posteriori del test globale
G = function(x,y)
{
  (lmbeta(x) - lmbeta(y))
}

test_prob = function(ph0,a,n,ny)
{
  lH0 = log(ph0) +  G(a + n,a)
  lH1 = log((1-ph0)) + sum(apply(ny,1,function(x) G(x + a,a)))
  out= exp(lH0)/( exp(lH0) + exp(lH1))
  return(out)
}
```

funzione per il calcolo delle probabilità delle classi latenti

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericMatrix get_prob_c(List psi, NumericMatrix nuh, IntegerVector y,
IntegerMatrix dati)
{
   NumericMatrix res(dati.nrow(),psi.size());
   NumericVector tmp_res(psi.size());
   List tmp_psi;
   NumericVector tmp_psi2;
   NumericVector tmp_nuh;
   double tmp_prob;

   for(int i =0; i <dati.nrow();i++)
   {
    // vettore pesi per l'individuo
    tmp_nuh = nuh(y(i),_);
    for(int h=0;h< psi.size();h++)
    {
      tmp_psi = psi(h);
      tmp_prob = 0;
      for(int j = 0; j< dati.ncol(); j++)
      {
        tmp_psi2 = tmp_psi(j);
        tmp_prob = tmp_prob + log(tmp_psi2(dati(i,j)));
       }
       res(i,h) = exp( log(tmp_nuh(h)) + tmp_prob);
     }

    res(i,_) =
          res(i,_)/std::accumulate(res(i,_).begin(),res(i,_).end(),0.0);
}

   return res;
}
```

Funzione di previsione

```
#include <Rcpp.h>
using namespace Rcpp;


// [[Rcpp::export]]
NumericMatrix previsione_individuo_y(
    List PSI ,
    ListOf<NumericMatrix> NUH,
    IntegerVector cat ,
    ListOf<NumericVector> PY,
    int start ,int H)
{
  List tmp_psi, tmp_psih;
    NumericMatrix tmp_nuh(H,4);
  NumericMatrix PREV(PSI.size()-start ,4);
  double tmp_prob;
  NumericVector prob_def(4),tmp_psihj;
  NumericVector tmp_py(4);
  NumericVector tmp_prob_h(H);
  NumericVector tmp_pred_n(H),tmp_pred_c(H),tmp_pred_s(H),tmp_pred_i(H);
  NumericVector tmp_nuh_n,tmp_nuh_c,tmp_nuh_s,tmp_nuh_i;
  for(int nsim=start; nsim< PSI.size();nsim ++)
  {
    tmp_psi   = PSI(nsim);
    tmp_nuh   = NUH[nsim];
    tmp_nuh_n = tmp_nuh(0,_);
    tmp_nuh_c = tmp_nuh(1,_);
    tmp_nuh_s = tmp_nuh(2,_);
    tmp_nuh_i = tmp_nuh(3,_);

    tmp_py = PY[nsim];
    for(int h=0; h <tmp_psi.size();h++)
    {
      tmp_psih = tmp_psi(h);
      tmp_prob = 0;

      for(int j=0; j<tmp_psih.size();j++)
      {
        tmp_psihj = tmp_psih(j);
        tmp_prob = tmp_prob + log(tmp_psihj(cat(j)));
      }

      tmp_pred_n(h) = exp(tmp_prob + log(tmp_nuh_n(h)));
      tmp_pred_c(h) = exp(tmp_prob + log(tmp_nuh_c(h)));
      tmp_pred_s(h) = exp(tmp_prob + log(tmp_nuh_s(h)));
      tmp_pred_i(h) = exp(tmp_prob + log(tmp_nuh_i(h)));
```

```
    }

    prob_def(0) = tmp_py(0) *
                  std::accumulate(tmp_pred_n.begin(),tmp_pred_n.end(),0.0);
    prob_def(1) = tmp_py(1) *
                  std::accumulate(tmp_pred_c.begin(),tmp_pred_c.end(),0.0);
    prob_def(2) = tmp_py(2) *
                  std::accumulate(tmp_pred_s.begin(),tmp_pred_s.end(),0.0);
    prob_def(3) = tmp_py(3) *
                  std::accumulate(tmp_pred_i.begin(),tmp_pred_i.end(),0.0);

    PREV(nsim - start,_) =
    prob_def/std::accumulate(prob_def.begin(),prob_def.end(),0.0);

  }
  return PREV;
}
```

Ringraziamenti

Innanzitutto desidero ringraziare il mio relatore, Professor Bruno Scarpa, sia per la sua disponibilità, i suoi consigli per la stesura di questa tesi, sia per avermi sempre spronato a dare il massimo in questi due anni ed aver contribuito a farmi appassionare sempre di più agli studi.
Desidero inoltre ringraziare il mio correlatore, Professor Giancarlo Ottaviano, per l'aiuto nelle parti "non statistiche" della tesi.
Uno speciale ringraziamento va a Daniele Durante per il prezioso aiuto e le "chiacchierate" sempre illuminanti.
Vorrei inoltre ringraziare tutti coloro che mi hanno dato una mano nel completamento di questo percorso: mio cugino Salvatore sempre pronto a consigliarmi e a correggere qualsiasi mio lavoro; Gabriella per la pazienza e per essere stata sempre disponibile; Tommaso, Caterina e Sebastiano per le nottate in aula studio; Sally per il cibo e per la compagnia. Infine voglio ringraziare la mia famiglia, mia madre, mio padre e mio fratello per avermi supportato e sopportato in questi anni e i miei amici di Padova e Benevento che sono davvero troppi da elencare ma che ringrazio con affetto per essere stati sempre vicino.

Bibliografia

Agresti, A. (2010). *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley.

Ayabe-Kanamura, S., Saito, S., Distel, H., MARTÍNEZ-GÓMEZ, M. & Hudson, R. (1998). Differences and similarities in the perception of everyday odors: A japanese-german cross-cultural study. *Annals of the New York Academy of Sciences* **855**, 694–700.

Azzalini, A. & Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. Oxford University Press.

Bensafi, M., Rinck, F., Schaal, B. & Rouby, C. (2007). Verbal cues modulate hedonic perception of odors in 5-year-old children as well as in adults. *Chemical senses* **32**, 855–862.

Bensafi, M., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M. & Holley, A. (2002). Autonomic nervous system responses to odours: the role of pleasantness and arousal. *Chemical Senses* **27**, 703–709.

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American statistical Association* **82**, 112–122.

Breiman, L. (2001). Random forests. *Machine learning* **45**, 5–32.

Capocasa, M., Anagnostou, P., Bachis, V., Battaggia, C., Bertoncini, S., Biondi, G., Boattini, A., Boschi, I., Brisighelli, F., Calò, C. M. et al. (2013). Linguistic, geographic and genetic isolation: a collaborative study of italian populations. *Journal of anthropological sciences= Rivista di antropologia: JASS/Istituto italiano di antropologia* **92**, 201–231.

Carroll, J. D. & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika* **35**, 283–319.

Croy, I., Olgun, S. & Joraschky, P. (2011). Basic emotions elicited by odors and pictures. *Emotion* **11**, 1331.

Distel, H., Ayabe-Kanamura, S., Martínez-Gómez, M., Schicker, I., Kobayakawa, T., Saito, S. & Hudson, R. (1999). Perception of everyday odors—correlation between intensity, familiarity and strength of hedonic judgement. *Chemical Senses* **24**, 191–199.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics* **6**, 241–252.

Dunson, D. B. & Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**, 1042–1051.

Durante, D. & Dunson, D. B. (2014). Bayesian inference on group differences in brain networks. *arXiv preprint arXiv:1411.6506* .

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211–218.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association* **90**, 577–588.

Friedman, J., Hastie, T. & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* **1**.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1.

Hitchcock, F. L. (1927). *The expression of a tensor or a polyadic as a sum of products.* sn.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

Hummel, T., Kobal, G., Gudziol, H. & Mackay-Sim, A. (2007). Normative data for the "sniffin'sticks" including tests of odor identification, odor discrimination, and olfactory thresholds: an upgrade based on a group of more than 3,000 subjects. *European Archives of Oto-Rhino-Laryngology* **264**, 237–243.

Ishwaran, H. & Zarepour, M. (2002a). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963.

Ishwaran, H. & Zarepour, M. (2002b). Exact and approximate sum representations for the dirichlet process. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* , 269–283.

Kolda, T. G. & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51**, 455–500.

Lazarsfeld, P. F., Henry, N. W. & Anderson, T. W. (1968). *Latent structure analysis.* Houghton Mifflin Boston.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review* **26**, 195–239.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 689–710.

Schab, F. R. (1990). Odors and the remembrance of things past. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**, 648.

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

Venables, W. N. & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®* **36**, 45–54.