Università degli studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Un algoritmo di topic modeling per microblog

Relatore: Prof. Emanuele Di Buccio
Dipartimento di Ingegneria dell'Informazione

Laureando: Giovanni Toto
Matricola N. 1242466

Anno Accademico 2021/2022

Indice

1 Introduzione **1**

1.1 Contributi della Tesi 3

1.2 Struttura della Tesi 4

2 Topic Modeling **7**

2.1 Modelli Generativi Probabilistici 8

2.2 Latent Dirichlet Allocation 9

2.2.1 Notazione 10

2.2.2 Specificazione del Modello 10

2.2.3 Processo Generativo 11

2.2.4 Distribuzione Congiunta 12

2.2.5 Inferenza 13

2.3 Origine ed evoluzione del Topic Modeling 13

2.3.1 Da pLSI a LDA 14

2.3.2 Estensioni della Latent Dirichlet Allocation 15

2.4 Topic Modeling per Microblog 18

2.4.1 Twitter-LDA 20

2.4.2 Hashtag-LDA 22

2.4.3 Dual-Sparse Topic Model 23

3 Modello Proposto **27**

3.1 Ipotesi ed Assunzioni 28

3.2 Notazione 35

3.3 Specificazione del Modello 36

3.3.1 Variabili e Distribuzioni a Livello di Collezione 36

3.3.2 Variabili e Distribuzioni a Livello di Utente 38

3.3.3 Variabili e Distribuzioni a Livello di Documento 38
3.3.4 Variabili e Distribuzioni a Livello di Parola 41
3.3.5 Variabili e Distribuzioni a Livello di Hashtag 43
3.3.6 Notazione Abbreviata 44
3.4 Processo Generativo 46
3.5 Derivazione della Distribuzione Congiunta 49
3.5.1 Caso di Partenza: LDA 49
3.5.2 Documenti che Trattano di un unico Topic e di più Topic 52
3.5.3 Parole di Sottofondo 57
3.5.4 Modello Finale: Parole e Hashtag 61
3.6 Relazione con Altri Modelli 67
3.6.1 Latent Dirichlet Allocation 68
3.6.2 Twitter-LDA 72
3.6.3 Hashtag-LDA 74

4 Inferenza tramite Collapsed Gibbs Sampler 77
4.1 Markov Chain Monte Carlo 78
4.1.1 Gibbs Sampler 79
4.1.2 Collapsed Gibbs Sampler 81
4.2 Costruzione del Collapsed Gibbs Sampler 81
4.2.1 Problema Inferenziale 82
4.2.2 Distribuzione Congiunta del Modello 82
4.2.3 Marginalizzazione dei Parametri 83
4.2.4 Full Conditional Probabilities 91
4.2.5 Stime dei Parametri 113
4.3 Blocked Collapsed Gibbs Sampler 121
4.4 Casi Particolari dell'Algoritmo 126
4.4.1 Latent Dirichlet Allocation 126
4.4.2 Twitter-LDA 127
4.4.3 Hashtag-LDA 128

5 Implementazione 131
5.1 Matrici dello Stato della Catena di Markov 133
5.2 Matrici di Conteggi 134
5.3 Inizializzazione dello Stato Iniziale e delle Matrici di Conteggi 140
5.4 Aggiornamento delle Variabili 140

5.5 Stime delle Variabili Latenti e dei Parametri 142

6 Esperimenti **145**

6.1 Costruzione della Collezione di Tweet 145

6.1.1 Download di Tweet tramite Twitter API 146

6.1.2 Pulizia dei Tweet 149

6.1.3 Filtro dei Tweet 150

6.2 Analisi Esplorative 150

6.3 Analisi Quantitative 154

6.3.1 Parametri 154

6.3.2 Convergenza 155

6.3.3 Metriche di Valutazione 156

6.4 Analisi Qualitative 163

6.4.1 Tipo dei Documenti 163

6.4.2 Rappresentazione dei Documenti come Mistura di Topic 165

6.4.3 Preferenze degli Utenti 165

6.4.4 Doppia Rappresentazione dei Topic 167

6.4.5 Parole di Sottofondo e Hashtag Globali 170

7 Conclusioni e Sviluppi Futuri **173**

A Distribuzioni di Probabilità **177**

A.1 Distribuzione Beta 177

A.2 Distribuzione di Dirichlet 177

A.3 Distribuzione di Bernoulli 178

A.4 Distribuzione Categoriale 178

A.5 Distribuzione Multinomiale 178

A.6 Confronto tra Distribuzione Categoriale e Distribuzione Multinomiale 179

B Logaritmo delle Distribuzioni Congiunte **181**

B.1 Latent Dirichlet Allocation 181

B.2 Twitter-LDA 181

B.3 Hashtag-LDA 182

B.4 Modello proposto 182

C Codice R **185**
C.1 Librerie 185
C.2 Collapsed Gibbs Sampler 186
C.3 Collapsed Gibbs Sampler (LDA) 200
C.4 Collapsed Gibbs Sampler (Twitter-LDA) 203
C.5 Collapsed Gibbs Sampler (Hashtag-LDA) 207
C.6 Metriche di Topic Coherence 212

D Codice Python **215**
D.1 Librerie 215
D.2 Download dei Tweet 215

Bibliografia **219**

Elenco delle figure

2.1 *Modello grafico probabilistico* della *Latent Dirichlet Allocation.* 12
2.2 *Modelli grafici probabilistici* di *Twitter-LDA* e *Hashtag-LDA*. . 21
2.3 *Modello grafico probabilistico* del *Dsparse TM*; la figura è tratta da Lin et al., 2014. 24

3.1 Divisione in blocchi del processo generativo. 29
3.2 Blocco 1: individuazione del tipo dei documenti. 30
3.3 Blocco 2.1: assegnazione dei topic ai documenti. 31
3.4 Blocco 2.0: assegnazione dei topic alle parole e agli *hashtag*. . 32
3.5 Blocco 3: generazione delle parole e degli *hashtag*. 33
3.6 *Modello grafico probabilistico* del modello proposto. 45
3.7 *Modello grafico probabilistico* della *LDA* come caso particolare del modello proposto. 69
3.8 *Modello grafico probabilistico* di *Twitter-LDA* come caso particolare del modello proposto. 72
3.9 *Modello grafico probabilistico* di *Hashtag-LDA* come caso particolare del modello proposto. 74

6.1 Distribuzione del numero di *tweet* pubblicati per ogni utente; le linee verticali sono i quartili della distribuzione. 151
6.2 Istogramma del numero di parole (sinistra) e del numero di *hashtag* (destra) contenuti in ogni *tweet*. 152
6.3 *Wordcloud* delle parole e degli *hashtag* contenuti nella collezione di *tweet*. 152
6.4 Lista delle 15 *top word* della distribuzione sulle parole, $\boldsymbol{\phi}^{\mathcal{C}}$, e lista dei 15 *top hashtag* della distribuzione sugli *hashtag*, $\boldsymbol{\psi}^{\mathcal{C}}$, della collezione di *tweet*. 153

6.5 *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i. 157
6.6 *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i. 158
6.7 *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i. 159
6.8 *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i. 160
6.9 Rappresentazione grafica di TC-PMI, TC-LCP, TC-NZ con $N = 10$ dei 16 *topic model* stimati; il grafico in basso a destra mostra la distanza media tra le righe di $\boldsymbol{\phi}_{1:T}$ e $\boldsymbol{\phi}^{\mathcal{C}}$, misurata con la *divergenza di Jensen-Shannon*. I *topic model* considerati sono: *Latent Dirichlet Allocation* (LDA), *Twitter-LDA* (TLDA), *Hashtag-LDA* (HLDA) e il modello proposto (MLDA). 162
6.10 Rappresentazione dei quattro documenti delle collezione che trattano di più topic come misture di topic; la linea orizzontale ha come ordinata $\frac{1}{30}$. 165
6.11 Distribuzione del numero di utenti per cui il topic considerato è quello preferito per ogni topic; l'indice k del topic preferito dell'utente u è dato da $k = \text{argmax}_{t \in \{1,\dots,T\}}\{\theta^*_{u,t}\}$. 166
6.12 Distribuzione del numero di utenti per cui il topic considerato è importante per ogni topic; un topic t è importante per l'utente u se $\theta_{u,t} \geq 0.2$. 167
6.13 Lista delle 15 *top word* e dei 15 *top hashtag* del topic 7. . . . 169
6.14 Lista delle 15 *top word* e dei 15 *top hashtag* del topic 23. . . . 169
6.15 Lista delle 15 *top word* e dei 15 *top hashtag* del topic 24. . . . 170
6.16 Lista delle 15 *top word* e dei 15 *top hashtag* del topic 29. . . . 170

6.17 Lista delle 15 *top word* e dei 15 *top hashtag* del topic 30. . . . 171
6.18 Lista delle 15 *top word* e dei 15 *top hashtag* del topic 22. . . . 171
6.19 Lista delle 15 *top word* della distribuzione sulle parole della collezione, $\boldsymbol{\phi}^{\mathbf{C}}$, e della distribuzione sulle parole delle parole di sottofondo, $\boldsymbol{\phi}^{\mathbf{B}}$. 172
6.20 Lista dei primi 15 *top hashtag* della distribuzione sugli *hashtag* della collezione, $\boldsymbol{\psi}^{\mathbf{C}}$, e della distribuzione sugli *hashtag* degli *hashtag globali*, $\boldsymbol{\psi}^{\mathbf{B}}$. 172

Elenco delle tabelle

3.1 Lista delle variabili osservate, delle variabili latenti e dei parametri del modello proposto; le variabili sono ordinate per livello, partendo dalla collezione fino ad arrivare alle parole e agli *hashtag*. L'indice u indica l'utente u, ud il d-mo documento scritto dall'utente u, udn l'n-ma parola del documento ud, udl l'l-mo *hashtag* del documento ud. 48

3.2 Lista delle variabili latenti presenti in *LDA*, *Twitter-LDA* e *Hashtag-LDA*. 68

5.1 Tabella dei conteggi; in basso si riporta il numero totale di conteggi contenuti nei vettori e nelle matrici. 139

6.1 Risultati delle analisi preliminari per la selezione del numero dei topic T; TC-PMI è calcolato su $N = 10$ *top word* di ogni topic. 155

6.2 TC-PMI, TC-LCP e TC-NZ con $N = 10$ dei 16 *topic model* stimati; l'ultima colonna contiene la distanza media tra le distribuzioni sulle parole dei topic e la distribuzione sulle parole della collezione, misurata con la *divergenza di Jensen-Shannon*. 164

Elenco degli algoritmi

1 *Gibbs Sampler* 80
2 *Collapsed Gibbs Sampler* 132
3 Inizializzazione dello stato iniziale e delle matrici di conteggi . 141
4 Aggiornamento di una variabile latente u_k 142
5 Stima delle variabili latenti e dei parametri 143

Capitolo 1

Introduzione

I *topic model* nascono come strumenti per ottenere una breve descrizione dei documenti di una collezione e allo stesso tempo preservare le relazioni statistiche essenziali tra le parole contenute nei testi; le rappresentazioni compatte di documenti sono utili, ad esempio, per effettuare classificazioni dei testi, costruire *filtri collaborativi*, valutare la similarità tra documenti o tra parole, valutare la rilevanza di documenti rispetto ad un'*interrogazione* (*query*) ad un motore di ricerca (Blei et al., 2003, Steyvers e Griffiths, 2007).

Questo problema di rappresentazione è affrontato dai *topic model* identificando all'interno di una collezione una serie di topic, definiti come distribuzioni di probabilità sulle parole di un vocabolario, e rappresentando i documenti come misture di quest'ultimi. Ad ogni topic corrisponde una tematica e i pesi che la distribuzione di ogni topic associa agli elementi del vocabolario possono essere utilizzati per identificarla. Ne consegue che la rappresentazione dei documenti come mistura di topic può facilitare l'accesso ed il reperimento dell'informazione permettendo, ad esempio, a utenti esperti –come ricercatori nell'ambito delle scienze sociali o giornalisti– di conoscere il contenuto –in termini di topic– di enormi collezioni di documenti senza dover necessariamente leggere ogni documento uno a uno.

I *topic model* –in particolare la *Latent Dirichlet Allocation*, che è il modello più rappresentativo– si sono dimostrati efficaci in diversi contesti applicativi, ad esempio quando utilizzati su collezioni di articoli di giornale e *abstract* accademici; tuttavia tendono a fornire risultati meno coerenti e interpretabili quando applicati ai post di *microblog*.

I *microblog* sono piattaforme virtuali che permettono ai loro utenti di comunicare attraverso la pubblicazione costante di piccoli contenuti, detti *micropost*, i quali possono contenere brevi messaggi di testo, immagini o video (Kaplan & Haenlein, 2011). Lo sviluppo dei *social media* –tra cui i *microblog*– nell'ultimo decennio ha spostato l'interesse di molti ricercatori verso l'estrazione di informazione da questa nuova forma di comunicazione caratterizzata da brevità, linguaggio informale –spesso con errori di battitura, acronimi e abbreviazioni non standard[1]–, utilizzo di *emoticon*, *emoji*, *tag*, *menzioni*, *hashtag* e molti altri elementi testuali non presenti nelle tipologie di testi su cui i primi *topic model* sono stati formulati (Mehrotra et al., 2013). Ad esempio, gli *hashtag* –sequenze di caratteri precedute dal simbolo #– assumono un ruolo molto importante all'interno di queste piattaforme dal momento che favoriscono la diffusione dell'informazione rendendo i *micropost* più facilmente reperibili e permettendo ad un utente di navigare più facilmente attraverso i contenuti della piattaforma. Questo linguaggio peculiare, ricco di sfumature in una lunghezza estremamente contenuta, rende difficile l'estrazione di informazione mediante tecniche originariamente formulate per testi molto più lunghi e con un linguaggio solitamente più formale e costante all'interno della collezione.

In questa tesi si considera il lavoro di W. X. Zhao et al., 2011, successivamente esteso in F. Zhao et al., 2016, in cui si abbandona una delle assunzioni principali del *topic modeling* –la rappresentazione di un documento come una mistura di topic– in favore di una rappresentazione semplificata secondo cui ogni documento tratta di un unico topic, i.e. di un'unica tematica. I modelli proposti nei due articoli, rispettivamente *Twitter-LDA* e *Hashtag-LDA*, creati ad-hoc per i *micropost* di *Twitter*, si basano sull'assunzione che, data la brevità dei testi considerati, tutti i documenti trattino di un'unica tematica e spostano la rappresentazione come mistura dei topic dai documenti agli autori[2] di essi. I modelli non forniscono più informazioni riguardanti i topic contenuti nei singoli documenti, ma forniscono informazioni sulle preferenze in termini di topic degli utenti considerati nella collezione. Inoltre nel secondo modello, *Hashtag-LDA*, è introdotta una distinzione tra parole

[1] Esempi di abbreviazioni non standard sono: *O o*, *ahahah*, *XD*, *LOL*, *YOLO*.

[2] Dal momento che l'autore di un *micropost* è un utente di una piattaforma, di seguito verrà utilizzato il termine *utente* al posto di autore.

e *hashtag* che permette di ottenere una doppia rappresentazione dei topic –solitamente espressi esclusivamente a partire dalle parole osservate– che permette un'analisi più ricca e approfondita delle tematiche della collezione.

1.1 Contributi della Tesi

L'assunzione secondo cui tutti i documenti trattano di un unico topic può funzionare se tutti i documenti della collezione sono brevi, tuttavia può diventare limitante quando la collezione è formata sia da documenti brevi e poco elaborati sia da documenti più lunghi e complessi. Lo scenario appena esposto è tipico di *piattaforme* come *Twitter* in cui la maggior parte dei *tweet* è formata da risposte semplici e concise alle pubblicazioni altrui, mentre una minoranza è formata da *tweet* più elaborati e complessi il cui scopo è esprimere un punto di vista originale, che potrebbe toccare più tematiche.

In questa tesi si riprende quindi il lavoro svolto in W. X. Zhao et al., 2011 e F. Zhao et al., 2016, proponendo un nuovo *topic model* che preserva tutti i pregi di *Twitter-LDA* e *Hahstag-LDA*, e allo stesso tempo cerca di alleviare l'assunzione sopra esposta, ritenuta troppo stringente. Più nello specifico, si ipotizza che considerare una distinzione tra documenti che trattano di un unico topic e documenti che trattano di più topic permetta di ottenere un "miglior" *topic model*, caratterizzato da topic significativi e facilmente interpretabili da un umano.

Il modello proposto è quindi costruito in modo tale da poter essere considerato un'estensione della *Latent Dirichlet Allocation*, di *Twitter-LDA* e *Hashtag-LDA*; in particolare, si riprende la struttura latente del primo *topic model* per gestire i documenti più complessi, che trattano di più topic, e si riprende la struttura latente degli altri due per gestire i documenti più semplici, che trattano di un unico topic. Per effettuare inferenza a posteriori approssimata, si propone un algoritmo *Collapsed Gibbs Sampler*; inoltre, si dimostra che i *Collapsed Gibbs Sampler* per l'inferenza della *LDA*, di *Twitter-LDA* e *Hashtag-LDA* possono essere facilmente ricavati a partire dalla formulazione dell'algoritmo del modello proposto in questa tesi. Questi quattro algoritmi sono stati quindi implementati in *R* e applicati ad una collezione di *tweet*, scaricata ad-hoc utilizzando le *Twitter API*. Il nuovo modello è stato confrontato quantitativamente con i suoi tre casi particolari –utilizzando metriche

proprie del *topic modeling*– e infine si è mostrato come estrarre informazioni dalla collezione di *tweet* interpretando la struttura latente del modello.

1.2 Struttura della Tesi

Il presente documento è articolato come segue:

Topic Modeling Nel Capitolo 2 si fornisce una definizione di *topic model* e la sua interpretazione come *modello generativo probabilistico*; si introduce quindi la *Latent Dirichlet Allocation* come modello di riferimento e come il *topic modeling* si è evoluto intorno ad essa; infine, si espongono i principali lavori che si sono concentrati su *microblog* o, più in generale, su testi brevi.

Modello Proposto Nel Capitolo 3 si introduce il modello proposto in questa tesi, evidenziando le ipotesi e le assunzioni su cui si basa; si specificano il suo *processo generativo* e il suo *modello grafico probabilistico*; si espone quindi come ottenere la distribuzione congiunta del modello proposto come estensione della *Latent Dirichlet Allocation*; infine, si mostra che *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA* possono essere considerati casi particolari del modello proposto in questa tesi.

Inferenza Tramite Collapsed Gibbs Sampler Nel Capitolo 4 si espone l'idea su cui si basano i metodi *Markov Chain Monte Carlo* e si introducono gli algoritmi *Gibbs Sampler* e *Collapsed Gibbs Sampler*; si procede quindi con la derivazione del *Collapsed Gibbs Sampler* per effettuare l'inferenza del modello proposto e si propone un *Blocked Collapsed Gibbs Sampler* come algoritmo alternativo; infine, si mostra come derivare i *Collapsed Gibbs Sampler* di *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA* a partire da quello del modello proposto.

Implementazione Nel Capitolo 5 si sposta l'attenzione dagli aspetti teorici alle accortezze necessarie per ottenere un algoritmo funzionante che permetta di ottenere delle stime attendibili in un periodo di tempo non eccessivamente lungo, ponendo particolare enfasi su come le varie quantità vengono salvate in memoria.

Esperimenti Nel Capitolo 6 si introduce *Twitter* e si mostra il procedimento utilizzato per ottenere un dataset strutturato su cui applicare il modello proposto e i suoi tre casi particolari; si propone quindi un possibile approccio per valutare e confrontare quantitativamente i quattro *topic model*; infine, si mostra come interpretare la struttura latente del modello proposto in questa tesi.

Capitolo 2

Topic Modeling

I *topic model* sono metodi statistici che, analizzando le parole contenute in una collezione di testi, permettono di scoprire le tematiche ricorrenti senza aver bisogno di ulteriori informazioni in aggiunta al testo stesso. Sono quindi metodi non supervisionati in grado di processare grandi moli di dati in maniera completamente automatica e ad una velocità impossibile per annotatori umani (Blei, 2012). Più nello specifico, i *topic model* sono *modelli mistura gerarchici* che –almeno nelle loro formulazioni più semplici– si basano sull'idea che ogni documento di una collezione sia modellato come una mistura di topic e che ogni topic sia caratterizzato da una distribuzione di probabilità sulle parole di un vocabolario noto e fissato (Steyvers & Griffiths, 2007).

Intuitivamente, a ogni topic corrisponde una tematica e il modello identifica una serie di topic, ma non la tematica ad essi associata: quest'ultima può essere identificata a partire dai pesi che la distribuzione di ogni topic associa agli elementi del vocabolario. Le parole con maggior peso –spesso dette *top word* in letteratura– possono essere utilizzate per associare un'etichetta al topic poiché, essendo le parole osservate più spesso, possono essere ritenute rappresentative della tematica del topic. Si consideri ad esempio un topic le cui prime 5 *top word* sono "pandemia", "covid19", "covid", "coronavirus" e "ospedale": è ragionevole assumere che la tematica associata al topic sia *Pandemia di COVID-19*.

Solitamente i *topic model* sono introdotti come *modelli generativi probabilistici* poiché quest'ultimi forniscono un modo molto semplice ed intuitivo

per esprimere il funzionamento del modello, che spesso presenta una struttura complessa e difficile da comprendere se espressa esclusivamente in termini probabilistici.

Nella prossima sezione si fornisce una definizione di *modelli generativi probabilistici* e si forniscono tre metodi per rappresentare uno stesso *topic model*; nella sezione 2.2 si introduce la *Latent Dirichlet Allocation* come modello di riferimento; nella sezione 2.3 si espongono l'origine dei *topic model* e le principali direzioni su cui si sono sviluppati; infine, nella sezione 2.4 si espongono i principali lavori che si sono concentrati su microblog o, più in generale, su testi brevi.

2.1 Modelli Generativi Probabilistici

Un *modello generativo probabilistico* per documenti testuali specifica una procedura statistica secondo cui dei documenti possono essere generati; le leggi statistiche che definiscono il modello si basano sull'utilizzo di variabili latenti, ovvero non osservabili. I dati –i termini nei testi dei documenti– sono essenzialmente trattati come se fossero generati a partire da un processo generativo a cui corrisponde una distribuzione congiunta di variabili latenti ed osservate (Blei, 2012). Ogni *modello generativo probabilistico* può essere descritto attraverso tre rappresentazioni equivalenti:

- distribuzione congiunta delle variabili latenti ed osservate;
- *processo generativo*;
- *modello grafico probabilistico*.

Distribuzione congiunta La distribuzione congiunta delle variabili latenti ed osservate è la rappresentazione più importante delle tre ed è fondamentale per poter definire procedure di inferenza, tuttavia è anche la rappresentazione meno intuitiva.

Processo generativo Il *processo generativo* di una collezione di documenti può essere definito come quel processo stocastico immaginario che si assume abbia generato i testi dei documenti (Srivastava & Sahami, 2009). Esso specifica la rappresentazione di ogni variabile aleatoria, da quale distribuzione è generata ed i legami che ha con le altre variabili del modello. Nel

caso specifico dei *topic model*, esiste una gerarchia tra variabili ed il processo generativo la esplora partendo dalle variabili definite a livello dell'intera collezione –i topic–, proseguendo con quelle a livello di documento –le proporzioni dei topic– fino ad arrivare a quelle a livello della singola parola. L'idea di fondo è che le caratteristiche generali della collezione –in questo caso i topic– influenzano i documenti contenuti in essa e, a loro volta, le caratteristiche di un documento influenzano il suo contenuto, ovvero il suo testo.

Modello grafico probabilistico Infine, un *modello grafico probabilistico* rappresenta un insieme di variabili aleatorie con le loro dipendenze condizionali attraverso l'uso di un grafo aciclico diretto. A ogni nodo corrisponde una variabile aleatoria in senso bayesiano –questa può essere una variabile osservata, una variabile latente oppure un parametro con una distribuzione a priori associata– e ogni arco rappresenta una relazione di dipendenza statistica tra due variabili. In particolare, la distribuzione di un nodo dipende esclusivamente dai nodi genitori e se due nodi non sono connessi, allora sono condizionalmente indipendenti tra loro dati i predecessori comuni. Replicazioni di una o più variabili sono rappresentate attraverso un rettangolo (*plate*) che riporta in basso a destra il numero di replicazioni: queste replicazioni si dicono *scambiabili*, ovvero la loro distribuzione congiunta è invariante a permutazioni.

2.2 Latent Dirichlet Allocation

La *Latent Dirichlet Allocation* (*LDA*) è un modello bayesiano composto da una gerarchia di modelli mistura in cui ogni documento è modellato come una mistura finita di topic in cui i pesi sono estratti una volta sola per ogni documento, ma i componenti della mistura –i topic– sono condivisi da tutti i documenti della collezione (Blei et al., 2010). L'intuizione dietro la struttura gerarchica dell'*LDA* può essere sintetizzata in quattro semplici punti:

- in una collezione di documenti esiste un numero fissato di topic; ogni topic è rappresentato come una distribuzione sui termini del vocabolario della collezione;

- ogni documento è rappresentato come una mistura di topic e il peso di un topic indica quanto è importante all'interno del documento;
- ogni parola ha un topic associato ed i topic con un peso maggiore all'interno di un documento sono assegnati più frequentemente alle sue parole;
- la parola effettivamente osservata dipende dal topic ad essa associato ed è più verosimile osservare gli elementi del vocabolario con un peso maggiore all'interno del topic.

Riepilogando, tutti i documenti trattano degli stessi topic, ma in proporzioni diverse; queste proporzioni influenzano le assegnazioni dei topic delle parole e quest'ultimi influenzano a loro volta le parole che vengono effettivamente osservate in un documento.

2.2.1 Notazione

Si consideri una collezione di D documenti, $\mathbf{w} = \{\mathbf{w}_1, \ldots, \mathbf{w}_d, \ldots, \mathbf{w}_D\}$ in cui il d-mo documento può essere rappresentato come una sequenza di N_d parole, $\mathbf{w}_d = \{w_{d1}, \ldots, w_{dn}, \ldots, w_{dN_d}\}$. Una parola è definita come una sequenza di caratteri a cui è assegnato un significato; una stessa sequenza può apparire in diversi documenti e tutte le V sequenze distinte osservate vanno a formare il vocabolario della collezione, indicizzato da $\{1, \ldots, V\}$. Formalmente, una parola $w_{dn} \in \{1, \ldots, V\}$ è uno scalare che assume il valore v se l'n-ma parola del d-mo documento è il v-mo elemento del vocabolario della collezione.

2.2.2 Specificazione del Modello

Si assume che i documenti della collezione trattino di un numero T fissato di topic. A ogni topic t è associata una distribuzione sui V elementi del vocabolario, $\boldsymbol{\phi}_t$, che segue una distribuzione di Dirichlet simmetrica di ordine V con parametro β^V, ovvero il topic t è rappresentato da un vettore di probabilità $V \times 1$ il cui v-mo elemento indica quanto è importante il v-mo elemento del vocabolario all'interno del topic t. Queste distribuzioni sono solitamente raccolte in una matrice $T \times V$ definita come segue

$$\boldsymbol{\phi}_{1:T} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \vdots \\ \boldsymbol{\phi}_T \end{bmatrix} = \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,V} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \cdots & \phi_{T,V} \end{bmatrix}$$

Analogamente a ogni documento d è associata una distribuzione sui T topic –detta *proporzioni dei topic*–, $\boldsymbol{\theta}_d$, che segue una distribuzione di Dirichlet simmetrica di ordine T con parametro α, ovvero il documento d è rappresentato da un vettore di probabilità $T \times 1$ il cui t-mo elemento indica quanto è importante il t-mo topic della collezione all'interno del documento d. Come sopra, le distribuzioni sono raccolte in una matrice $D \times T$ definita come segue

$$\boldsymbol{\theta}_{1:D} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_D \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \dots & \theta_{1,T} \\ \vdots & \ddots & \vdots \\ \theta_{D,1} & \dots & \theta_{D,T} \end{bmatrix}$$

Gli elementi delle matrici $\boldsymbol{\phi}_{1:T}$ e $\boldsymbol{\theta}_{1:D}$ sono i parametri delle distribuzioni delle parole e dei topic ad esse associati; le distribuzioni di Dirichlet sono quindi le distribuzioni a priori dei parametri del modello.

Per ogni parola dn è introdotta una variabile latente z^V_{dn} che indica il topic associato alla parola; si usa la stessa rappresentazione utilizzata per le parole, quindi $z^V_{dn} \in \{1, \dots, T\}$ è uno scalare che assume il valore t se il topic associato all'n-ma parola del d-mo documento è il t-mo topic della collezione. Riprendendo la notazione utilizzata per $\mathbf{w}_d$, i topic assegnati alle parole del d-mo documento possono essere indicati con $\mathbf{z}^V_d = \{z^V_{d1}, \dots, z^V_{dn}, \dots, z^V_{dN_d}\}$. La distribuzione condizionata del topic associato all'n-ma parola del d-mo documento, z^V_{dn}, date le proporzioni dei topic del documento a cui appartiene, $\boldsymbol{\theta}_d$, segue una distribuzione categoriale con vettore di probabilità $\boldsymbol{\theta}_d$:

$$z^V_{dn} | \boldsymbol{\theta}_d \sim Cat(\boldsymbol{\theta}_d)$$

Infine, la distribuzione condizionata dell'n-ma parola del d-mo documento, w_{dn}, dati il topic ad essa associato, z^V_{dn}, e le distribuzioni sulle parole[1] dei topic della collezione, $\boldsymbol{\phi}_{1:T}$, segue una distribuzione categoriale con vettore di probabilità $\boldsymbol{\phi}_{z^V_{dn}}$:

$$w_{dn} | z^V_{dn}, \boldsymbol{\phi}_{1:T} \sim Cat(\boldsymbol{\phi}_{z^V_{dn}})$$

2.2.3 Processo Generativo

Il processo generativo della *Latent Dirichlet Allocation*, rappresentato in Figura 2.1, è il seguente:

[1] Per compattezza, d'ora in poi si scriverà semplicemente *distribuzioni sulle parole* al posto di *distribuzione sui V elementi del vocabolario.*

1. Per ogni topic $t = 1, \dots, T$:
   a. Si estrae la distribuzione sulle parole del topic t da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\phi}_t|\beta^V \sim Dir_V(\beta^V)$.
2. Per ogni documento $d = 1, \dots, D$:
   a. Si estrae la distribuzione sui topic del documento d da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\theta}_d|\alpha \sim Dir_T(\alpha)$.
   b. Per ogni parola $n = 1, \dots, N_d$:
      i. Si estrae il topic della parola dn dalla distribuzione sui topic del documento d, $z^V_{dn}|\boldsymbol{\theta}_d \sim Cat(\boldsymbol{\theta}_d)$.
      ii. Si estrae la parola dn dalla distribuzione sulle parole del topic ad essa associato, $w_{dn}|z^V_{dn}, \boldsymbol{\phi}_{1:T} \sim Cat(\boldsymbol{\phi}_{z^V_{dn}})$.

Queste due rappresentazioni equivalenti della *Latent Dirichlet Allocation* forniscono un modo molto semplice e intuitivo per capire come le variabili latenti[2] ed osservate sono legate tra loro.

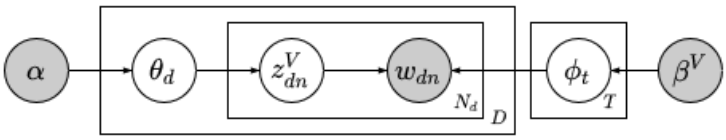


Figura 2.1: *Modello grafico probabilistico* della *Latent Dirichlet Allocation*.

2.2.4 Distribuzione Congiunta

Assumendo che l'ordine dei documenti all'interno della collezione –assunzione di scambiabilità dei documenti– e l'ordine delle parole all'interno di un documento –assunzione di scambiabilità delle parole o assunzione *bag-of-words*– non siano rilevanti, la distribuzione congiunta delle variabili osservate e latenti dati i parametri fissati α e β^V è data da:

$$
\begin{aligned}
& p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}|\alpha, \beta^V) \\
&= \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\beta^V) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N_d} p(z^V_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z^V_{dn}, \boldsymbol{\phi}_{1:T}) \\
&= \left(\prod_{t=1}^{T} \frac{\Gamma(V\beta^V)}{\prod_{v=1}^{V} \Gamma(\beta^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta^V - 1} \right) \prod_{d=1}^{D} \left(\frac{\Gamma(T\alpha)}{\prod_{t=1}^{T} \Gamma(\alpha)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha - 1} \right)
\end{aligned}
$$

[2]Si noti che anche i parametri contenuti in $\boldsymbol{\phi}_{1:T}$ e $\boldsymbol{\theta}_{1:D}$ sono variabili latenti.

$$\times \prod_{n=1}^{N_d} \left(\prod_{t=1}^{T} \theta_{d,t}^{\mathbb{1}_{z_{dn}^V = t}} \prod_{v=1}^{V} \phi_{z_{dn}^V, v}^{\mathbb{1}_{w_{dn}=v}} \right)$$

2.2.5 Inferenza

L'obiettivo della *Latent Dirichlet Allocation* è stimare la struttura latente definita dal modello, ovvero stimare tutto ciò che è legato ai topic: le distribuzioni sui topic dei documenti contenute in $\boldsymbol{\theta}_{1:D}$, le distribuzioni sulle parole dei topic contenute in $\boldsymbol{\phi}_{1:T}$ e i topic associati alle parole contenuti in $\mathbf{z}^V$. Il problema inferenziale che deve essere risolto è quindi il calcolo della distribuzione posteriori delle variabili latenti date le variabili osservate:

$$p(\mathbf{z}^V, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T} | \mathbf{w}, \alpha, \beta^V) = \frac{p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T} | \alpha, \beta^V)}{p(\mathbf{w} | \alpha, \beta^V)}$$

Il denominatore $p(\mathbf{w}|\alpha, \beta^V)$ è ottenuto marginalizzando le variabili latenti

$$p(\mathbf{w}|\alpha, \beta^V) = \sum_{\mathbf{z}^V} \int_{S_{\boldsymbol{\theta}_{1:D}}} \int_{S_{\boldsymbol{\phi}_{1:T}}} p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T} | \alpha, \beta^V) d\boldsymbol{\theta}_{1:D} d\boldsymbol{\phi}_{1:T}$$

dove l'integrale rispetto a $\boldsymbol{\theta}_{1:D}$ corrisponde a D integrali sul simplesso $(T-1)$-dimensionale e l'integrale rispetto a $\boldsymbol{\phi}_{1:T}$ corrisponde a T integrali sul simplesso $(V-1)$-dimensionale. A causa di questi integrali non è possibile calcolare il denominatore e di conseguenza non è possibile effettuare inferenza esatta sulla struttura latente del modello.

Diversi algoritmi sono stati proposti nell'arco degli anni per effettuare inferenza a posteriori approssimata, tra cui *Variational Inference* (Blei et al., 2003), *Expectation-Maximization algorithm* (Blei et al., 2003), e *Gibbs Sampling* (Griffiths & Steyvers, 2004).

2.3 Origine ed evoluzione del Topic Modeling

Rappresentare in maniera strutturata il contenuto di un testo è un aspetto critico di ogni approccio di *Information Retrieval*, o più in generale di ogni approccio di *Text Mining*, e l'utilizzo di una rappresentazione piuttosto che un'altra può influenzare enormemente la quantità e la qualità delle informazioni estraibili da una collezione di documenti testuali.

Uno dei primi approcci che ha associato una serie di valori numerici a un documento è il *Bag-of-Words* (BoW) *model* in cui l'ordine delle parole non

è rilevante e ogni documento è rappresentato da un vettore $V \times 1$ il cui v-mo elemento è il numero di occorrenze del v-mo elemento del vocabolario della collezione. Alternativamente, il numero di occorrenze –*term frequency* (*tf*)– può essere sostituito con il *term frequency-inverse document frequency* (*tf-idf*) in modo da ridurre il peso delle parole che compaiono frequentemente in tutti i documenti della collezione. L'approccio è interessante poiché permette di identificare le parole più importanti, tuttavia la dimensione elevata dei vettori fornisce poche informazioni riguardo la struttura statistica interna a ogni documento e tra documenti (Blei et al., 2003).

I *topic model* nascono come un metodo alternativo per ottenere una rappresentazione vettoriale più compatta rispetto alla precedente; in particolare, ogni documento è rappresentato come un vettore $T \times 1$, dove T è il numero di topic: solitamente questo numero è molto più piccolo rispetto all'ampiezza V del vocabolario. Due *topic model* fondamentali sono la *Latent Dirichlet Allocation* e il *probabilistic Latent Semantic Indexing* (*pLSI*, Hofmann, 1999), anche noto come *aspect model*: il primo è di fatto il *topic model* per eccellenza, essendo sia il più famoso sia quello maggiormente utilizzato come base di partenza per costruire nuovi modelli, mentre il secondo può essere considerato il predecessore della *LDA* che ha costruito le fondamenta per la nascita di quest'ultima.

2.3.1 Da pLSI a LDA

pLSI è il primo modello che considera un documento come una mistura di topic e nasce per rilassare l'assunzione forte della *mistura di unigrammi* (*mixture of unigrams*, Nigam et al., 2000) secondo cui tutte le parole di un documento hanno associato lo stesso topic. *pLSI* ha essenzialmente lo stesso processo generativo della *LDA*, tuttavia non associa una distribuzione a priori né alle distribuzioni sui topic dei documenti né alle distribuzioni sulle parole dei topic. L'assenza di un processo generativo probabilistico a livello di documento crea due seri problemi: (1) il numero di parametri del modello cresce linearmente con la dimensione della collezione, portando a seri problemi di sovradattamento; (2) non è chiaro come assegnare le proporzioni dei topic a un documento non presente nella collezione utilizzata per stimare i parametri del modello.

La versione originale della *LDA* è un'estensione di *pLSI* in cui viene introdotta una distribuzione a priori per le proporzioni dei topic dei documenti: il modello così ottenuto è a tutti gli effetti un *modello generativo probabilistico* che non presenta più i due problemi sopra esposti. Si noti che il modello esposto nella sezione 2.2 in realtà è lo *smoothed LDA*, ovvero un'estensione completamente bayesiana della *LDA* in cui è assegnata una distribuzione a priori anche ai topic della collezione. *LDA* e *smoothed LDA* sono entrambi proposti in Blei et al., 2003, tuttavia il secondo è maggiormente approfondito nell'articolo in cui viene proposto un *Collapsed Gibbs Sampler* come algoritmo per effettuare l'inferenza (Griffiths & Steyvers, 2004). È importante fare questa distinzione poiché il modello che in letteratura viene solitamente chiamato *LDA* è in realtà lo *smoothed LDA*.

2.3.2 Estensioni della Latent Dirichlet Allocation

Una caratteristica estremamente interessante della *LDA* –che ha fatto la sua fortuna– è la modularità che permette di formulare in maniera relativamente semplice estensioni del modello originale aggiungendo nuove variabili latenti, modificando le distribuzioni di probabilità già presenti nel modello, o facendo entrambe le cose.

Rilassamento delle assunzioni

Nei primi lavori successivi a quello di Blei et al., 2003, i ricercatori si concentrano sulle tre assunzioni di base della *LDA*:

- numero di topic fissato,
- scambiabilità dei documenti,
- scambiabilità delle parole.

Ritenendole in alcuni casi troppo stringenti, propongono estensioni che rilassano tali assunzioni.

La scelta del numero di topic T può influenzare fortemente l'interpretabilità dei risultati: una soluzione con T troppo basso tende a portare topic molto vaghi, mentre una con T troppo alto tende a portare topic difficilmente interpretabili. Si hanno essenzialmente tre possibilità per la selezione del numero di topic: utilizzare una conoscenza a priori del dataset, stimare lo stesso modello con diversi T e selezionarne uno attraverso un criterio

fissato,[3] ricorrere a metodi non parametrici in cui il modello seleziona autonomamente il numero appropriato di topic (Blei et al., 2010). In particolare, Teh et al., 2006 propongono un metodo bayesiano non parametrico, detto *Hierarchical Dirichlet Process* (*HDP*), che permette di ottenere un *topic model* in cui il numero dei topic diventa parte della distribuzione a posteriori dalla struttura latente. Per costruire il nuovo modello, le proporzioni dei topic estratte da una Dirichlet simmetrica sono sostituite da distribuzioni sui topic G_d estratte da un *processo di Dirichlet* (*DP*); la distribuzione di base di questi *DP* –uno per ogni documento– è a sua volta estratta da un *DP*. Questo approccio è molto interessante poiché può essere facilmente integrato nella maggior parte dei *topic model* parametrici, convertendoli in approcci non parametrici, senza complicare eccessivamente la procedura d'inferenza.

L'assunzione di scambiabilità dei documenti, ovvero l'assunzione secondo cui l'ordine temporale dei documenti non è rilevante, può essere ragionevole per collezioni che raccolgono documenti scritti tutti circa lo stesso periodo, tuttavia è irrealistica se si considera una collezione formata da testi scritti in momenti temporali molto differenti, ad esempio a decenni di distanza. Due dei primi lavori che rilassano l'assunzione di scambiabilità dei documenti sono il *Dynamic Topic Model* (*DTM*, Blei e Lafferty, 2006) e *Topic Over Time* (*TOT*, Wang e McCallum, 2006). Il primo assume che la collezione sia organizzata in epoche, ogni epoca sia caratterizzata dai suoi topic e che ognuno di essi dipenda dallo stesso topic all'epoca precedente; si assume quindi che i documenti siano scambiabili solo all'interno della stessa epoca. A livello pratico, per la prima epoca si stima la *LDA*, mentre dalla seconda in poi un modello quasi identico in cui la distribuzione sulle parole di ogni topic è modellata tramite un *modello state-space con rumore gaussiano*. Per evitare sia di discretizzare il tempo sia di considerare processi Markoviani su cui è difficile fare inferenza, il secondo utilizza un approccio differente in cui la collocazione nel tempo dei documenti e le co-occorrenze delle parole sono modellati congiuntamente

L'assunzione della scambiabilità delle parole, o assunzione *bag-of-words*, è ragionevole per motivi computazionali ed inferenziali, di conseguenza è l'aspetto meno approfondito in letteratura. Tuttavia, qui di seguito, si introduce brevemente un *topic model* il cui processo generativo è stato utilizzato

[3] I problema del confronto di *topic model* è trattato nella sottosezione 6.3.3.

come riferimento per la costruzione del modello proposto. Il modello introdotto in Griffiths et al., 2005 considera sia le dipendenze sintattiche tra parole vicine all'interno di una frase sia le dipendenze semantiche tra parole anche lontane all'interno di uno stesso documento; in particolare, viene utilizzato un *modello composito* (*composite model*) che permette di gestire il fatto che tutte le parole presentano dipendenze sintattiche, ma solo alcune anche dipendenze semantiche. Ad ogni parola sono assegnati un topic e una classe: l'assegnazione dei topic avviene come nella *LDA*, mentre l'appartenenza a una determinata classe dipende dalla classe della parola precedente attraverso un *Hidden Markov Model* (*HMM*). Ad ogni topic e ad ogni classe è associata una distribuzione sugli elementi del vocabolario: una parola viene estratta dalla distribuzione del topic ad essa associato se appartiene a una particolare classe, detta *classe semantica*; altrimenti, viene estratta dalla distribuzione della classe ad essa associata. L'aspetto interessante del modello è che tutte le parole hanno sia un topic sia una classe associata, nel senso che vengono estratti per tutte le parole, tuttavia il topic ha significato per una determinata parola solo se essa appartiene alla *classe semantica*.

Utilizzo di metadati

Parallelamente, altri ricercatori si concentrano su come incorporare metadati nel modello. Il primo modello a sfruttare informazioni oltre al testo è l'*author-topic model* (Rosen-Zvi et al., 2004) in cui le proporzioni sui topic non sono più associate ai documenti, ma agli autori di quest'ultimi; in particolare, per ogni parola un autore è estratto casualmente tra quelli associati al documento, quindi un topic viene estratto dalla proporzione dei topic propria di quell'autore e infine una parola è estratta dalla distribuzione sulle parole propria di quel topic. Più in generale, una qualsiasi informazione aggiuntiva disponibile –autore, data di pubblicazione, titolo del documento, locazione geografica, presenza di link, *hashtag*, ...– è legata a un documento attraverso il *tagging*. Un *tag* è associato a un documento se quest'ultimo presenta la caratteristica associata al tag: ad esempio, se il tag "Autore: David M. Blei" è associato a un documento, ciò significa che David M. Blei è l'autore del documento; se "Pubblicazione: 22/02/2022" è associato, ciò significa che 22/02/2022 è la data in cui il documento è stato pubblicato. Sfruttando questa intuizione, nasce il *tag-topic model* (Tsai, 2011), un modello che coin-

cide essenzialmente con l'*author-topic model*, ma, al posto di determinare le distribuzioni sui topic degli autori, determina le distribuzioni sui topic dei *tag* assegnati ai *blog*. Un'ulteriore estensione è data da *Tag-Latent Dirichlet Allocation* (*TLDA*, Ma et al., 2013) in cui al posto di selezionare casualmente un *tag* tra quelli osservati, il modello assegna a ogni documento una proporzione dei *tag* e utilizza quest'ultima per estrarre il *tag* assegnato a ogni parola.

Si noti che questi modelli assumono che siano i *tag* a influenzare la presenza o meno di determinati topic in un documento: questo approccio nasce dall'idea che il *tag* sia un elemento slegato dal testo. Al contrario, il modello proposto in questa tesi si basa esattamente sull'opposto, ovvero i *tag* –in questo caso specifico gli *hashtag*– sono parte del testo e come tali sono loro ad essere influenzati dai topic e non viceversa; questa assunzione è ripresa da *Hashtag-LDA* (F. Zhao et al., 2016), un modello che modella congiuntamente parole e *hashtag* che verrà introdotto nella prossima sezione.

Ulteriori approfondimenti

In questa sezione sono stati descritti brevemente il contesto del *topic modeling*, il ragionamento che porta la creazione di nuovi modelli e di come molti di questi non siano altro che leggere variazione di lavori precedenti. Questo approccio alla disciplina ha ispirato il modello proposto in questa tesi: esso presenta elementi innovativi e allo stesso tempo elementi propri di lavori precedenti. Si rimanda infine all'articolo di Jelodar et al., 2019 per una trattazione più dettagliata dello sviluppo del *topic modeling* avvenuto tra il 2003 e il 2016.

2.4 Topic Modeling per Microblog

La sviluppo dei *social media* nell'ultimo decennio ha spostato l'interesse di molti ricercatori verso l'estrazione di informazione dai *microblog* pubblicati dagli utenti di queste piattaforme; in particolare, ciò ha portato a sviluppi di nuovi approcci per la gestione e l'analisi di testi strutturalmente molto diversi, in termini sia di lunghezza sia di contenuto, rispetto a quelli studiati in precedenza.

Hong e Davison, 2010 analizzano il problema della modellazione di testi brevi (*short text modeling*) applicando la *LDA* e l'*author-topic model* a diverse collezioni in cui i documenti sono aggregazioni di *tweet* e confrontano i risultati con quelli ottenuti su una collezione in cui nessuna operazione di aggregazione è stata applicata. Il lavoro non propone nuovi modelli, ma evidenzia le limitazioni dei *topic model* tradizionali quando applicati a contesti come quello dei *social media* in cui testi molto lunghi sono sostituiti da post di *microblog*.

Lin et al., 2014 osservano che contenuti generati dagli utenti nei *social media* sono caratterizzati da una estrema brevità, un ampio vocabolario e di conseguenza anche un grande numero di topic; dato il numero di topic, ritengono ragionevole pensare che ogni contenuto tratti solo di un piccolo gruppo di topic e analogamente ognuno di questi topic utilizzi un numero ristretto di parole tra tutte quelle disponibili. Nasce quindi il *Dual-sparse Topic Model* (*Dsparse TM*), un modello che gestisce sia la sparsità nelle proporzioni dei topic dei documenti sia la sparsità nell'utilizzo delle parole all'interno di ogni topic.[4]

Infine, due modelli creati ad-hoc per i *tweet* di *Twitter* sono sono *Twitter-LDA* (W. X. Zhao et al., 2011) e *Hashtag-LDA* (F. Zhao et al., 2016). Il primo ha come scopo principale identificare i topic trattati dagli utenti della piattaforma e confrontarli con quelli estratti dai *media tradizionali*, mentre il secondo nasce come metodo di raccomandazione di *hashtag*. L'assunzione che distingue questi due metodi da quasi tutti gli altri *topic model* è che viene associato un unico topic a ogni *tweet*; più nello specifico, ogni parola in *Twitter-LDA* e ogni *hashtag* in *Hashtag-LDA* può derivare dall'unico topic del *tweet* di appartenenza oppure da una distribuzione di sottofondo comune a tutta la collezione. Il secondo modello può essere visto come un'estensione del primo in cui gli *hashtag* contenuti nei *tweet* sono generati come le parole nel primo, cioè dalla distribuzione di uno dei topic o da una distribuzione comune. È interessante notare che F. Zhao et al., 2016 non considerano una distribuzione delle parole di sottofondo e assumono quindi che tutte le parole di un *tweet* siano associate a un topic: questa assunzione probabilmente è stata fatta per semplificare il modello –velocizzare l'inferenza– e poiché

[4]Il modello è detto *Dual-sparse* poiché gestisce la doppia sparsità (*dual-sparsity*) nelle distribuzioni dei documenti e dei topic.

l'interesse primario è sviluppare un metodo di raccomandazione, non ottenere topic interpretabili.

Nelle prossime sezioni si introducono brevemente *Twitter-LDA*, *Hashtag-LDA* e *Dsparse TM*: i primi due poiché sono stati utilizzati come base di partenza del modello proposto e possono essere considerati come suoi casi particolari; il terzo poiché il suo approccio per la gestione della sparsità dei topic a livello di documento è stato rielaborato ed incluso nel modello proposto.

2.4.1 Twitter-LDA

Twitter-LDA (W. X. Zhao et al., 2011) è un modello creato ad-hoc per i *tweet* di *Twitter* in cui si assume che ogni utente abbia una sua proporzione dei topic, $\boldsymbol{\theta}_u^*$, ogni *tweet* abbia un unico topic, z_{ud}^*, e ogni parola possa essere generata a partire da un topic oppure essere di sottofondo (*background words*): nel primo caso, si estrae la parola dalla distribuzione sulle parole del topic, $\boldsymbol{\phi}_t$; nel secondo caso, si estrae una parola dalla distribuzione sulle parole delle parole di sottofondo, $\boldsymbol{\phi}^{\mathcal{B}}$. Il processo generativo di *Twitter-LDA*, rappresentato in Figura 2.2a, è il seguente:

1. Per ogni topic $t = 1, \ldots, T$:
    a. Si estrae la distribuzione sulle parole del topic t da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\phi}_t | \boldsymbol{\beta}^V \sim Dir_V(\beta^V)$.
2. Si estrae la distribuzione sulle parole delle parole di sottofondo da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\phi}^{\mathcal{B}} | \beta^V \sim Dir_V(\beta^V)$.
3. Si estrae la probabilità di avere una parola generata a partire da un topic da una distribuzione di Dirichlet simmetrica[5], $\pi^V | b^V \sim Dir_2(b^V)$.
4. Per ogni utente $u = 1, \ldots, U$:
    a. Si estrae la distribuzione sui topic dell'utente u da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\theta}_u^* | \alpha^* \sim Dir_T(\alpha^*)$.
    b. Per ogni documento $d = 1, \ldots, D_u$:
        i. Si estrae il topic del documento ud dalla distribuzione sui topic dell'utente u, $z_{ud}^* \sim Cat(\boldsymbol{\theta}_u^*)$.

[5]Si noti che una distribuzione di Dirichlet simmetrica di ordine 2 con parametro b^V coincide con una distribuzione Beta con parametri uguali e pari a b^V, $Dir_2(b^V) \equiv Beta(b^V, b^V)$.

ii. Per ogni parola $n = 1, \ldots, N_{ud}$:

A. Si estrae l'origine della parola *udn* da una distribuzione di Bernoulli, $y^V_{udn}|\pi^V \sim Bern(\pi^V)$.

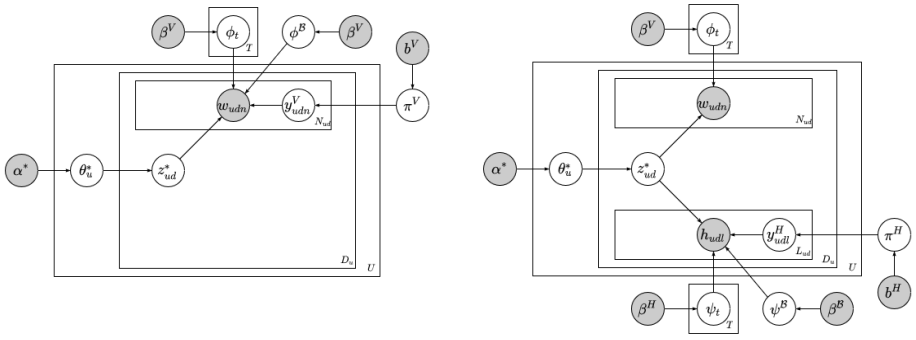
B. Se $y^V_{udn} = 0$, si estrae la parola *udn* dalla distribuzione sulle parole delle parole di sottofondo,

$$w_{udn}|y^V_{udn} = 0, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}} \sim Cat(\boldsymbol{\phi}^{\mathcal{B}});$$

se $y^V_{udn} = 1$, si estrae la parola *udn* dalla distribuzione sulle parole del topic associato al documento *ud*,

$$w_{udn}|y^V_{udn} = 1, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}} \sim Cat(\boldsymbol{\phi}_{z^*_{ud}}).$$

Le proporzioni dei topic, $\boldsymbol{\theta}^*_u$, forniscono informazioni su quali topic l'utente u tende a scrivere e l'analisi delle proporzioni sui topic di tutti gli utenti può essere utile sia per determinare gli interessi dei singoli utenti sia per identificare gli argomenti di tendenza dell'intera collezione analizzata. La distribuzione sulle parole delle parole di sottofondo, $\boldsymbol{\phi}^{\mathcal{B}}$, gestisce la presenza di parole molto utilizzate all'interno della collezione che, senza questa distribuzione, risulterebbero importanti all'interno della maggior parte dei topic; la distinzione tra parole di sottofondo e parole generate a partire da un topic permette quindi di ottenere topic di maggior qualità, ovvero più facilmente interpretabili e con diverse *top word*.



(a) *Twitter-LDA* **(b)** *Hashtag-LDA*

Figura 2.2: *Modelli grafici probabilistici* di *Twitter-LDA* e *Hashtag-LDA*.

2.4.2 Hashtag-LDA

Hashtag-LDA (F. Zhao et al., 2016) è un modello basato sulla *LDA* che modella congiuntamente la relazione tra utenti, *hashtag* e parole nei *microblog* attraverso dei topic latenti. Ad ogni post è associato un unico topic, z^*_{ud}, estratto dalle proporzioni dei topic del suo autore, $\boldsymbol{\theta}^*_u$, e ad ogni topic sono associate due distribuzioni, una sulle parole, $\boldsymbol{\phi}_t$, e una sugli *hashtag*, $\boldsymbol{\psi}_t$. Da queste ultime si generano separatamente le parole e gli *hashtag* contenuti nei *microblog*: le prime sono generate come nella *LDA*, i secondi possono essere generati dalla distribuzione sugli *hashtag* del topic associato al post oppure dalla distribuzione degli *hashtag globali*, $\boldsymbol{\psi}^{\mathcal{B}}$, che è comune a tutti i topic e a tutti gli autori. Il processo generativo di *Hashtag-LDA*, rappresentato in Figura 2.2b, è il seguente:

1. Per ogni topic $t = 1, \dots, T$:
    a. Si estrae la distribuzione sulle parole del topic t da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\phi}_t|\beta^V \sim Dir_V(\beta^V)$.
    b. Si estrae la distribuzione sugli *hashtag* del topic t da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\psi}_t|\beta^H \sim Dir_H(\beta^H)$.
2. Si estrae la distribuzione sugli *hashtag* degli *hashtag globali* da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\psi}^{\mathcal{B}}|\beta^{\mathcal{B}} \sim Dir_H(\beta^{\mathcal{B}})$.
3. Si estrae la probabilità di avere un *hashtag* generato a partire da un topic da una distribuzione di Dirichlet simmetrica, $\pi^H|b^H \sim Dir_2(b^H)$.
4. Per ogni utente $u = 1, \dots, U$:
    a. Si estrae la distribuzione sui topic dell'utente u da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\theta}^*_u|\alpha^* \sim Dir_T(\alpha^*)$.
    b. Per ogni documento $d = 1, \dots, D_u$:
        i. Si estrae il topic del documento ud dalla distribuzione sui topic dell'utente u, $z^*_{ud} \sim Cat(\boldsymbol{\theta}^*_u)$.
        ii. Per ogni parola $n = 1, \dots, N_{ud}$:
            A. Si estrae la parola udn dalla distribuzione sulle parole del topic associato al documento ud, $w_{udn}|\boldsymbol{\phi}_{1:T}, z^*_{ud} \sim Cat(\boldsymbol{\phi}_{z^*_{ud}})$.
        iii. Per ogni *hashtag* $l = 1, \dots, L_{ud}$:

A. Si estrae l'origine della parola *udl* da una distribuzione di Bernoulli, $y^H_{udl}|\pi^H \sim Bern(\pi^H)$.

B. Se $y^H_{udl} = 0$, si estrae l'*hashtag udl* dalla distribuzione sugli *hashtag* degli *hashtag globali*,

$$h_{udl}|y^H_{udl} = 0, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}} \sim Cat(\boldsymbol{\psi}^{\mathcal{B}});$$

se $y^H_{udl} = 1$, si estrae l'*hashtag udl* dalla distribuzione sugli *hashtag* del topic associato al documento *ud*,

$$h_{udl}|y^H_{udl} = 1, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}} \sim Cat(\boldsymbol{\psi}_{z^*_{ud}}).$$

L'interpretazione di $\boldsymbol{\theta}^*_u$ è la stessa di *Twitter-LDA*, mentre la distribuzione sugli *hashtag* degli *hashtag globali*, $\boldsymbol{\psi}^{\mathcal{B}}$, gestisce la presenza di *hashtag* molto utilizzati all'interno della collezione. Per ogni topic t si hanno due distribuzioni, una per le parole, $\boldsymbol{\phi}_t$, e una per gli hashtag, $\boldsymbol{\psi}_t$, che rappresentano la stessa tematica attraverso due vocabolari diversi. Solitamente le due distribuzioni sono molto simili ed entrambe interpretabili, tuttavia potrebbe essere necessario guardare entrambe poiché una delle due distribuzioni non risulta abbastanza informativa per determinare la *label*[6] del topic.

2.4.3 Dual-Sparse Topic Model

Dual-sparse Topic Model (*DsparseTM*, Lin et al., 2014) è un modello che utilizza un *processo Spike and Slab* per gestire sia la sparsità nelle proporzioni dei topic dei documenti sia la sparsità nell'utilizzo delle parole all'interno di ogni topic. Se un documento tratta di pochi topic, oppure se un topic è caratterizzato da pochi termini del vocabolario, allora è detto sparso.

Per ogni topic k è determinato un insieme di termini attivi, detto insieme dei *focused term* e indicato con B_k, che influenza la distribuzione a priori sulle parole del topic k; in particolare, siano $\overline{\gamma}$ e γ due parametri fissati, detti *weak term smoothing prior* e *term smoothing prior*, allora la distribuzione sui topic del topic k, $\vec{\phi}_k$, segue una distribuzione di Dirichlet di parametro

$$\gamma \vec{\beta}_k + \overline{\gamma} \vec{1} = \gamma \begin{bmatrix} \beta_{k1} \\ \vdots \\ \beta_{kV} \end{bmatrix} + \overline{\gamma} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

[6] Per *label* si intende una o più parole che riassumono la tematica trattata dal topic.

dove β_{kv} è una variabile dicotomica latente, detta *term selector*, che indica se il v-mo termine del vocabolario è un *focused term* per il topic k. Assumendo $\overline{\gamma} \ll \gamma$, si ha che la *weak term smoothing prior* $\overline{\gamma}$ definisce la distribuzione a priori dei termini non selezionati dal *term selector*, mentre la *term smoothing prior* definisce la distribuzione a priori dei termini selezionati, visto che $\overline{\gamma}$ risulta trascurabile nella somma $\gamma + \overline{\gamma}$. Fissando inoltre $\overline{\gamma} \to 0$, si ha che gli elementi della distribuzione sulle parole del topic k, $\vec{\phi}_k$, corrispondenti ai valori non selezionati assumono valori talmente bassi –ma non nulli– da poter essere assunti pari a zero: è ragionevole quindi affermare che $\sum_{r \in B_k} \phi_{kr} = 1$. Grazie a quest'ultima proprietà, il vettore $\{\vec{\phi}_{kr} : r \in B_k\}$ di dimensione $|B_k| \times 1$ può essere considerato un vettore di probabilità e utilizzato come vettore di probabilità delle distribuzioni categoriali delle parole.

Lo stesso ragionamento vale per i *focused topic*: l'idea chiave di *DsparseTM* è quindi utilizzare queste distribuzioni a priori per ridurre la dimensione sia del simplesso delle parole sia del simplesso dei topic nelle distribuzioni di Dirichlet al fine di introdurre sparsità nel modello.

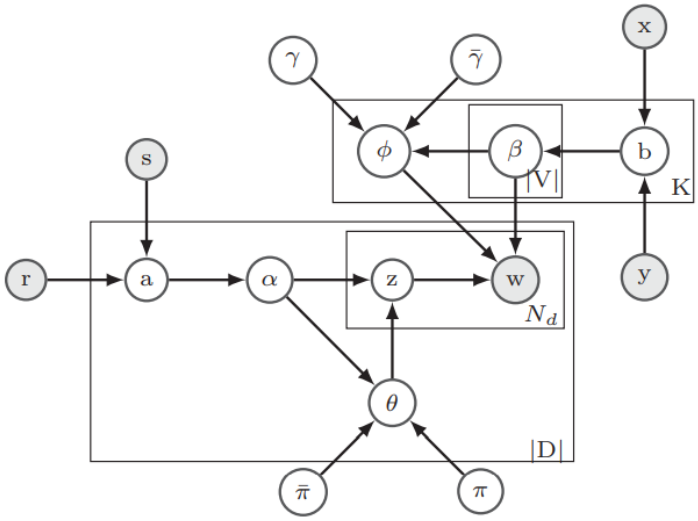


Figura 2.3: *Modello grafico probabilistico* del *Dsparse TM*; la figura è tratta da Lin et al., 2014.

Infine, il processo generativo di *DsparseTM*, rappresentato in Figura 2.3, è il seguente:

1. Per ogni topic $k = 1, \ldots, K$:
   a. Si estrae il parametro della distribuzione di Bernoulli associata al topic k, $b_k \sim Beta(x, y)$.
   b. Per ogni termine $r = 1, \ldots, V$:
      i. Si estrae il *term selector*, $\beta_{kr} \sim Bern(b_k)$ con $\vec{\beta}_k = \{\beta_{kr}\}_{r=1}^{V}$.
      ii. Si crea l'insieme dei *focused term*, $B_k = \{r : \beta_{kr} = 1\}$.
      iii. Si estrae la distribuzione sulle parole da una distribuzione di Dirichlet in cui si assegna una probabilità a priori maggiore ai *focused term*, $\vec{\phi}_k \sim Dir_V(\gamma\ \vec{\beta}_k + \overline{\gamma}\ \vec{1})$.
2. Per ogni documento $d = 1, \ldots, D$:
   a. Si estrae il parametro della distribuzione di Bernoulli associata al documento d, $a_d \sim Beta(s, t)$.
   b. Per ogni topic $k = 1, \ldots, K$:
      i. Si estrae il *topic selector*, $\alpha_{dk} \sim Bern(a_d)$ con $\vec{\alpha}_d = \{\alpha_{dk}\}_{k=1}^{K}$.
      ii. Si crea l'insieme dei *focused topic*, $A_d = \{k : \alpha_{dk} = 1\}$.
   c. Si estrae la distribuzione sui topic del documento d da una distribuzione di Dirichlet in cui si assegna una probabilità a priori maggiore ai *focused topic*, $\vec{\theta}_d \sim Dir_K(\pi\ \vec{\alpha}_d + \overline{\pi}\ \vec{1})$.
   d. Per ogni parola $i = 1, \ldots, N_d$:
      i. Si estrae il topic della parola *di* dalla distribuzione sui topic del documento d, $z_{di} \sim Mult(\{\vec{\theta}_{dk} : k \in A_d\})$.
      ii. Si estrae la parola *di* dalla distribuzione sulle parole del topic ad essa associato, $w_{di} \sim Mult(\{\vec{\phi}_{z_{di}r} : r \in B_{z_{di}}\})$.

Capitolo 3

Modello Proposto

Studi precedenti (F. Zhao et al., 2016; W. X. Zhao et al., 2011) suggeriscono che associare un unico topic a ogni post di *microblog*, o *micropost*, sia un'assunzione ragionevole poiché la maggior parte dei *tweet* –*micropost* della piattaforma di *microblogging* *Twitter*– tratta di un'unica tematica. Lin et al., 2014 propongono un *topic model* che associa un peso non trascurabile solo a un sottogruppo dei topic in ogni documento e osservano che il modello ottiene risultati leggermente inferiori –in termini di *TC-PMI*[1]– rispetto a un modello che associa un unico topic quando viene applicato ai *tweet*.

In questa tesi si ipotizza quindi che la maggior parte dei documenti tratti di un unico topic e una piccola parte di due o più; in particolare, si ipotizza che considerare nel processo generativo questa distinzione tra documenti che trattano di un unico topic e documenti che trattano di più topic permetta di ottenere un "miglior" *topic model*, ovvero permetta di identificare topic di maggior qualità –più significativi, coerenti e facilmente interpretabili– e permetta di assegnare quest'ultimi ai documenti e agli utenti in maniera utile e appropriata.

L'idea è considerare un modello molto generale che considera sia parole sia *hashtag*, sia documenti che trattano di un unico topic sia documenti che trattano di più topic, sia distribuzioni dei topic sia distribuzioni di sottofondo per la modellazione dei testi di una collezione di *micropost*. Il processo generativo è costruito in modo tale da poter essere considerato un'estensione

[1] *TC-PMI* è una metrica per misurare la *topic coherence* di un *topic model* basata sul *Pointwise Mutual Information*; questa ed altre metriche verranno introdotte nella sottosezione 6.3.3.

della *LDA*, di *Twitter-LDA* e *Hashtag-LDA*. In particolare, si riprende la struttura latente del primo *topic model* per gestire i documenti più complessi che trattano di più topic e si riprende la struttura latente degli altri due per gestire i documenti più semplici che trattano di un unico topic.

Nella sezione 3.1 si specificano l'ipotesi da verificare –esistenza di documenti che trattano di un unico topic e documenti che trattano di più topic– e le assunzioni su cui si basa il modello proposto; nelle sezione 3.2 si introduce la notazione e nella sezione 3.3 si definiscono le distribuzioni delle variabili aleatorie presenti nel modello; nella sezione 3.4 si riportano il *processo generativo* e il *modello grafico probabilistico*; nella sezione 3.5 si espone come ottenere la distribuzione congiunta del modello proposto come estensione della *Latent Dirichlet Allocation*; infine, nella sezione 3.6 si mostra che *LDA*, *Twitter-LDA* e *Hashtag-LDA* possono essere considerati casi particolari del modello proposto.

3.1 Ipotesi ed Assunzioni

Si riprendono le assunzioni di *LDA*, *Twitter-LDA* e *Hashtag-LDA* e se ne aggiungono di ulteriori per verificare l'ipotesi che considerare nel processo generativo una distinzione tra documenti che trattano di un unico topic e documenti che trattano di più topic permetta di ottenere un "miglior" *topic model*. Data la maggiore complessità, in questo caso si preferisce descrivere il modello in termini di processo generativo e dividere quest'ultimo in blocchi concettuali per spiegare in maniera più semplice possibile l'intuizione dietro la sua struttura gerarchica. In particolare, il processo generativo del modello proposto può essere diviso nei tre blocchi riportati in Figura 3.1: nel blocco 1 si ha l'individuazione del tipo dei documenti, nel blocco 2 si ha l'assegnazione dei topic ai documenti, alle parole e agli *hashtag* della collezione, nel blocco 3 si ha la generazione delle parole e degli *hashtag* osservati nei testi. Si noti che il blocco due può essere a sua volta diviso in due blocchi, nominati 2.1 e 2.0.

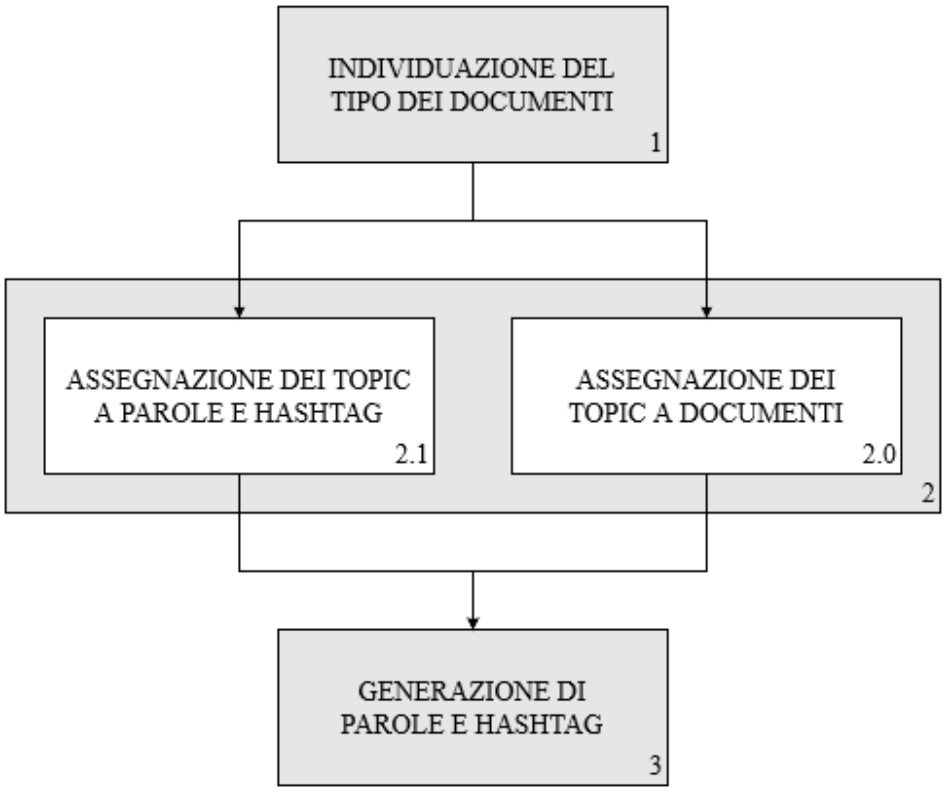


Figura 3.1: Divisione in blocchi del processo generativo.

Di seguito si analizzano in maniera discorsiva i tre blocchi sopra introdotti senza addentrasi negli aspetti più teorici; quest'ultimi verranno trattati nella sezione successiva. Per facilitare ulteriormente la comprensione della struttura del modello, per ogni blocco si riporta un modello grafico probabilistico in cui sono evidenziate solo le variabili legate ad esso.

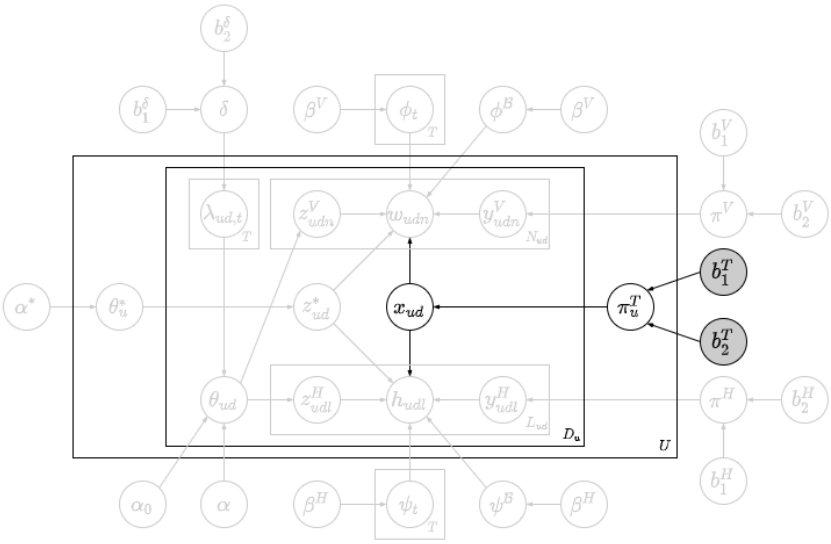


Figura 3.2: Blocco 1: individuazione del tipo dei documenti.

Blocco 1: Individuazione del tipo dei documenti

Nel primo blocco, rappresentato in Figura 3.2, è introdotta l'idea chiave su cui si basa l'intero modello: l'esistenza di due tipi di documenti, quelli che trattano di un unico topic e quelli che trattano di più topic. Si ipotizza che ogni utente u abbia una sua tendenza, rappresentata da una probabilità π_u^T, a scrivere documenti che trattano di un unico topic o di più topic e tale tendenza influenza il tipo x_{ud} dei documenti scritti dall'utente. Intuitivamente, diversi utenti possono utilizzare in maniera differente i *social media* e di conseguenza scrivere *micropost*[2] strutturalmente anche molto diversi tra loro: alcuni utenti –ad esempio pagine ufficiali di personalità politiche o del mondo dello spettacolo, *influencers*– scrivono solitamente *micropost* molto lunghi ed elaborati poiché utilizzano la piattaforma per esprimere il loro punto di vista, altri utenti si limitano semplicemente a rispondere alle pubblicazioni altrui attraverso *micropost* concisi e semplici.

[2]Si utilizza il termine *micropost* quando si vuole spiegare il contesto; si utilizza il termine più generale *documento* quando si descrive il modello.

Blocco 2: Assegnazione dei topic

Nel secondo blocco avviene l'assegnazione dei topic ai documenti, alle parole e agli *hashtag* della collezione; in particolare, l'assegnazione può avvenire in due modi in base ai tipi identificati nel primo blocco: se un documento tratta di più topic, l'assegnazione dei topic avviene in maniera molto simile alla *LDA*, ovvero ad ogni parola e ad ogni *hashtag* è associato un topic; se un documento tratta di un unico topic, l'assegnazione dei topic avviene come in *Twitter-LDA* e *Hashtag-LDA*, ovvero a ogni documento è associato un topic, detto *topic principale* per distinguerlo dai topic associati alle parole e agli *hashtag*.

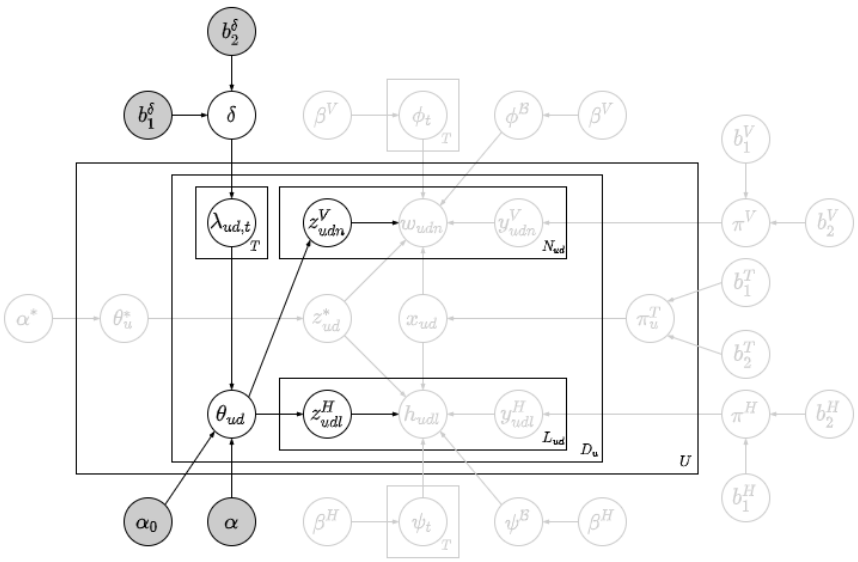


Figura 3.3: Blocco 2.1: assegnazione dei topic ai documenti.

Documenti che trattano di più topic Nel blocco 2.1, rappresentato in Figura 3.2, avviene l'assegnazione dei topic alle parole e agli *hashtag* della collezione in un processo generativo molto simile a quello della *LDA*:

- a ogni documento *ud* è associata una proporzione dei topic, $\boldsymbol{\theta}_{ud}$, il cui t-mo elemento indica quanto il t-mo topic della collezione è importante all'interno del documento;
- a ogni documento *ud* è associato un vettore dei topic attivi, $\boldsymbol{\lambda}_{ud}$, che influenza la proporzione dei topic associata al documento: se un topic

non è attivo, allora il suo peso è talmente basso da rendere essenzialmente impossibile osservare parole e *hashtag* con quel topic associato all'interno del documento; l'attivazione di un topic in un documento è guidata da una probabilità, δ, comune all'intera collezione;

- a ogni parola *udn* è associato un topic, z^V_{udn}, ed i topic con un peso maggiore all'interno di un documento sono assegnati più frequentemente alle sue parole; analogamente, a ogni *hashtag udl* è associato un topic, z^H_{udl}, ed i topic con un peso maggiore all'interno di un documento sono assegnati più frequentemente ai suoi *hashtag*.

Intuitivamente, si assume che anche i *micropost* più lunghi ed elaborati non contengano un numero troppo elevato di topic e che si focalizzino solo su un numero ristretto di essi: si considera quindi vettore dei topic attivi per introdurre sparsità nella rappresentazione come misture di topic dei documenti.

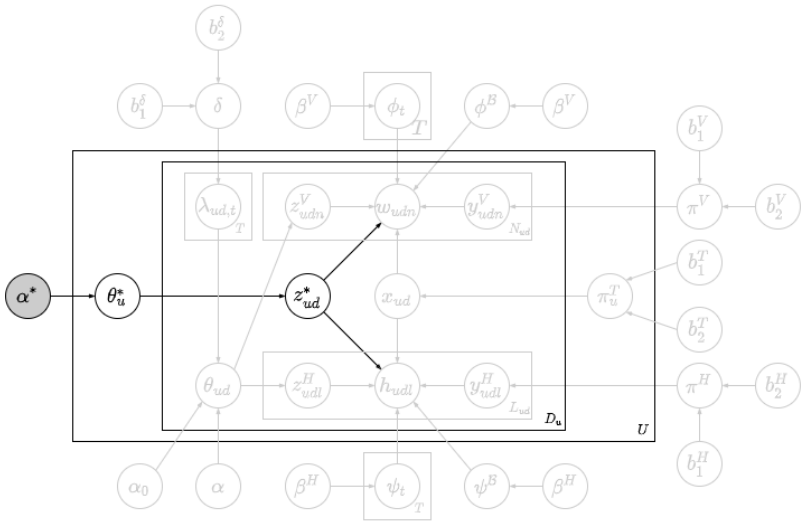


Figura 3.4: Blocco 2.0: assegnazione dei topic alle parole e agli *hashtag*.

Documenti che trattano di un unico topic Nel blocco 2.0, rappresentato in Figura 3.4, avviene l'assegnazione dei topic ai documenti della collezione in un processo generativo uguale a quello di *Twitter-LDA* e *Hashtag-LDA*:

- a ogni utente u è associata una proporzione dei topic, $\boldsymbol{\theta}^*_u$, il cui t-mo elemento indica quanto l'utente tende a scegliere il t-mo topic come topic principale di un documento;
- a ogni documento ud è associato un topic principale, z^*_{ud}, ed i topic con un peso maggiore per un utente sono assegnati più frequentemente ai suoi documenti come topic principali;
- alle parole e agli *hashtag* non è associato un topic poiché si assume che tutti gli elementi del testo generati a partire da un topic siano stati generati a partire dal topic principale del documento in cui sono contenuti.

Intuitivamente, si assume che i *micropost* più semplici e concisi trattino di un'unica tematica e che la scelta di quest'ultima dipenda esclusivamente dai gusti personali dell'utente che l'ha scritto in quanto spesso i *micropost* di questo tipo sono scritti d'impulso in risposta a pubblicazioni d'interesse. Per modellare la presenza di un'unica tematica, si abbandona la rappresentazione di un documento come mistura di topic e si opta per assegnare i topic ai documenti e non più agli elementi del testo.

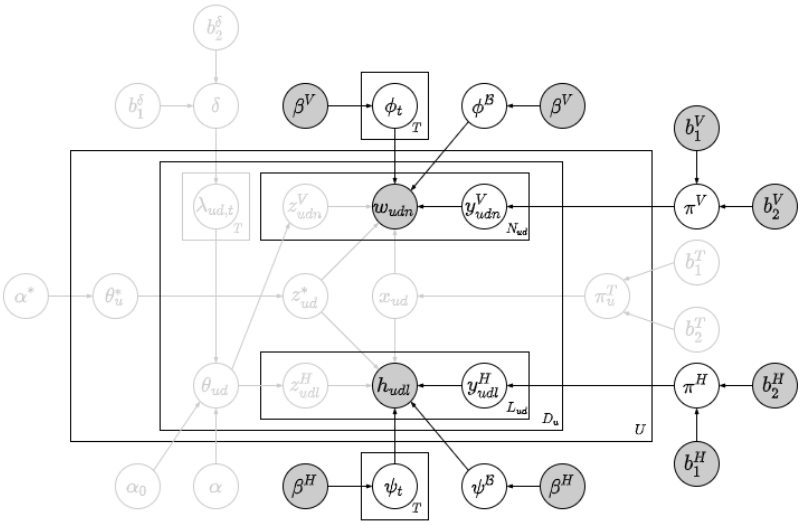


Figura 3.5: Blocco 3: generazione delle parole e degli *hashtag*.

Blocco 3: Generazione delle parole e degli hashtag

Nel terzo e ultimo blocco, rappresentato in Figura 3.5, avviene la generazione delle parole e degli *hashtag* osservati a partire dai topic trattati nei documenti

della collezione. Prima di tutto si introducono la doppia rappresentazione dei topic e la presenza di parole di sottofondo e *hashtag globali*:

- in una collezione di documenti esiste un numero fissato di topic; ogni topic è rappresentato sia come una distribuzione sulle parole del vocabolario sia come una distribuzione sugli *hashtag* osservati nella collezione;
- in una collezione di documenti esistono parole di sottofondo che sono comuni a tutti i topic: questo gruppo di parole è trattato come un topic ed è quindi rappresentato come una distribuzione sulle parole del vocabolario; analogamente, esistono *hashtag globali* che sono molto diffusi e utilizzati nei documenti indipendentemente dal topic di cui trattano: questo gruppo di *hashtag* è trattato come un topic ed è quindi rappresentato come una distribuzione sugli *hashtag* osservati nella collezione.

Infine, si riporta come avviene la distinzione tra parole generate a partire da un topic e parole di sottofondo, tra *hashtag* generati a partire da un topic e *hashtag globali*, e come le parole e gli *hashtag* effettivamente osservati nel testo vengono generati sulla base della loro origine, dei topic ad essi associati e del tipo dei documenti di appartenenza:

- a ogni parola *udl* è associata un'origine, y^V_{udn}, che indica se è stata generata a partire da un topic oppure se è una parola di sottofondo; analogamente, a ogni *hashtag udl* è associata un'origine, y^H_{udl}, che indica se è stato generato a partire da un topic oppure se è un *hashtag globale*;
- la parola effettivamente osservata, w_{udn}, dipende dalla sua origine, dal topic ad essa associato e dal tipo del documento di appartenenza: se la parola è di sottofondo si considera la distribuzione sulle parole delle parole di sottofondo, $\boldsymbol{\phi}^{\mathcal{B}}$, altrimenti si considera la distribuzione del topic principale, $\boldsymbol{\phi}_{z^*_{ud}}$ o quella del topic associato alla parola, $\boldsymbol{\phi}_{z^V_{udn}}$, a seconda che il documento tratti di uno o più topic; analogamente, l'*hashtag* effettivamente osservato dipende dalla sua origine, dal topic ad esso associato e dal tipo del documento di appartenenza: se l'*hashtag* è globale si considera la distribuzione sugli *hashtag* degli *hashtag globali*, $\boldsymbol{\psi}^{\mathcal{B}}$, altrimenti si considera la distribuzione del topic principale, $\boldsymbol{\psi}_{z^*_{ud}}$ o

quella del topic associato all'*hashtag*, $\boldsymbol{\psi}_{z^H_{udl}}$, a seconda che il documento tratti di uno o più topic.

Si noti che i processi generativi delle parole e degli *hashtag* sono essenzialmente identici, cambiano solo i parametri considerati: per ogni topic t si hanno due distribuzioni, una sugli elementi del vocabolario delle parole, $\boldsymbol{\phi}_t$, e una sugli elementi del vocabolario degli *hashtag*, $\boldsymbol{\psi}_t$, che possono essere utilizzate per generare le parole e gli *hashtag* osservati nei documenti. Distinguere parole e *hashtag* significa assumere che gli elementi del testo di un documento non provengano tutti dalla stesso vocabolario, ma da vocabolari diversi definiti secondo un determinato criterio; sebbene in questo caso si considerino parole e *hashtag*, l'approccio può essere generalizzato, includendo ad esempio altri vocabolari, come le *emoji*.

3.2 Notazione

Si consideri una collezione di $D = \sum_{u=1}^{U} D_u$ documenti scritti da U utenti, dove D_u è il numero di documenti scritti dall'utente u; a ogni documento è associato un unico utente. Sotto l'assunzione *bag-of-words*, il d-mo *tweet* scritto dall'utente u può essere rappresentato come una sequenza di N_{ud} parole $\mathbf{w}_{ud} = \{w_{ud1}, \ldots, w_{udn}, \ldots, w_{udN_{ud}}\}$ e una sequenza di L_{ud} *hashtag* $\mathbf{h}_{ud} = \{h_{ud1}, \ldots, h_{udl}, \ldots, h_{udL_{ud}}\}$; sia per le parole sia per gli *hashtag* non è rilevante l'ordinamento all'interno del testo. Una parola è definita come una sequenza di caratteri a cui è assegnato un significato; una stessa sequenza può apparire in diversi documenti e tutte le V sequenze distinte osservate vanno a formare il vocabolario delle parole della collezione, indicizzato da $\{1, \ldots, V\}$. Formalmente, una parola $w_{udn} \in \{1, \ldots, V\}$ è uno scalare che assume il valore v se l'n-ma parola del d-mo documento scritto dall'utente u è il v-mo elemento del vocabolario delle parole della collezione. Un *hashtag* è definito come una sequenza di caratteri preceduta dal simbolo #; una sequenza di caratteri preceduta dal simbolo # può apparire in diversi documenti e tutte le H sequenze distinte osservate vanno a formare il vocabolario degli *hashtag* della collezione, indicizzato da $\{1, \ldots, H\}$. Formalmente, un *hashtag* $h_{udl} \in \{1, \ldots, H\}$ è uno scalare che assume il valore h se l'h-mo *hashtag* del d-mo documento scritto dall'utente u è l'h-mo elemento del vocabolario degli *hashtag* della collezione.

Per comodità, di seguito, si indicherà con ud il d-mo documento scritto dall'utente u; analogamente si indicheranno rispettivamente con udn e udl la n-ma parola del d-mo documento scritto dall'utente u e l'h-mo *hashtag* del d-mo documento scritto dall'utente u.

3.3 Specificazione del Modello

In questa sezione si definiscono le variabili latenti e le distribuzioni di probabilità di tutte le variabili aleatorie del modello per livelli: prima si introducono le quantità comuni a tutti i documenti di tutti gli utenti della collezione, poi si introducono le quantità legate ai singoli utenti e ai singoli documenti, infine quelle legate alle singole parole e ai singoli *hashtag*.

In Tabella 3.1 sono riportate le variabili aleatorie e le loro distribuzioni. Si noti che le variabili aleatorie a livello di parola e a livello di *hashtag* corrispondenti hanno le stesse distribuzioni di probabilità con parametri differenti: ad esempio, y^V_{udn} e y^H_{udl} seguono entrambe delle distribuzioni di Bernoulli, ma la prima con parametro π^V, mentre la seconda con parametro π^H.

3.3.1 Variabili e Distribuzioni a Livello di Collezione

Si assume che i documenti della collezione trattino di un numero T fissato di topic. A ogni topic t sono associate:

- una distribuzione sui V elementi del vocabolario delle parole, $\boldsymbol{\phi}_t$, che segue una distribuzione di Dirichlet di ordine V con parametro $\boldsymbol{\beta}^V$, ovvero il topic t è rappresentato da un vettore di probabilità $V \times 1$ il cui v-mo elemento indica quanto è importante il v-mo elemento del vocabolario delle parole all'interno del topic t,
- una distribuzione sugli H elementi del vocabolario delle *hashtag*, $\boldsymbol{\psi}_t$, che segue una distribuzione di Dirichlet di ordine H con parametro $\boldsymbol{\beta}^H$, ovvero il topic t è rappresentato da un vettore di probabilità $H \times 1$ il cui h-mo elemento indica quanto è importante l'h-mo elemento del vocabolario degli *hashtag* all'interno del topic t.

Le distribuzioni relative alle parole sono raccolte in una matrice $T \times V$ definita come segue

$$\boldsymbol{\phi}_{1:T} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \vdots \\ \boldsymbol{\phi}_T \end{bmatrix} = \begin{bmatrix} \phi_{1,1} & \dots & \phi_{1,V} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \dots & \phi_{T,V} \end{bmatrix}$$

mentre quelle relative agli *hashtag* sono raccolte in una matrice $T \times H$ definita come segue

$$\boldsymbol{\psi}_{1:T} = \begin{bmatrix} \boldsymbol{\psi}_1 \\ \vdots \\ \boldsymbol{\psi}_T \end{bmatrix} = \begin{bmatrix} \psi_{1,1} & \dots & \psi_{1,H} \\ \vdots & \ddots & \vdots \\ \psi_{T,1} & \dots & \psi_{T,H} \end{bmatrix}$$

Alle parole di sottofondo è associata una distribuzione sui V elementi del vocabolario delle parole, $\boldsymbol{\phi}^{\mathcal{B}}$, che segue una distribuzione di Dirichlet di ordine V con parametro $\boldsymbol{\beta}^V$, ovvero le parole di sottofondo sono rappresentate da un vettore di probabilità $V \times 1$ il cui v-mo elemento indica quanto è probabile osservare il v-mo elemento del vocabolario delle parole quando si sta considerando una parola di sottofondo. Analogamente, agli *hashtag globali* è associata una distribuzione sugli H elementi del vocabolario degli *hashtag*, $\boldsymbol{\psi}^{\mathcal{B}}$, che segue una distribuzione di Dirichlet di ordine H con parametro $\boldsymbol{\beta}^H$, ovvero gli *hashtag globali* sono rappresentati da un vettore di probabilità $H \times 1$ il cui h-mo elemento indica quanto è probabile osservare l'h-mo elemento del vocabolario degli *hashtag* quando si sta considerando un *hashtag globale*.

La probabilità che una parola sia generata a partire da un topic, π^V, segue una distribuzione Beta con parametri b_1^V e b_2^V; essa rappresenta la tendenza generale degli utenti a utilizzare parole appartenenti a una specifica tematica e non parole più generali, ovvero le parole di sottofondo. Analogamente, la probabilità che un *hashtag* sia generato a partire da un topic, π^H, segue una distribuzione Beta con parametri b_1^H e b_2^H; essa rappresenta la tendenza generale degli utenti a utilizzare *hashtag* appartenenti a una specifica tematica e non *hashtag* più generali, ovvero gli *hashtag globali*.

La probabilità che un topic sia attivo in un documento, δ, segue una distribuzione Beta con parametri b_1^δ e b_2^δ; essa rappresenta la tendenza ad avere molte tematiche trattate nei documenti della collezione quando questi ultimi sono rappresentati come misture di topic.

Gli elementi delle matrici $\boldsymbol{\phi}_{1:T}$, $\boldsymbol{\psi}_{1:T}$, $\boldsymbol{\psi}^{\mathcal{B}}$, $\boldsymbol{\psi}^{\mathcal{B}}$ e le probabilità π^V, π^H e δ sono i parametri delle distribuzioni delle parole, degli *hashtag*, della presenza di un topic in un documento e dell'origine di parole ed *hashtag*; queste distribuzioni sono quindi le distribuzioni a priori delle variabili $\mathbf{w}$, $\mathbf{h}$, $\mathbf{y}^V$, $\mathbf{y}^H$ e $\boldsymbol{\lambda}$.

3.3.2 Variabili e Distribuzioni a Livello di Utente

A ogni utente u è associata una distribuzione sui T topic, $\boldsymbol{\theta}^*_u$, che segue una distribuzione di Dirichlet di ordine T con parametro $\boldsymbol{\alpha}^*$, ovvero l'utente u è rappresentato da un vettore di probabilità $T \times 1$ il cui t-mo elemento indica la tendenza dall'utente u a scegliere il topic t come topic principale. Queste distribuzioni sono raccolte in una matrice $U \times T$ definita come segue

$$\boldsymbol{\theta}^*_{1:U} = \begin{bmatrix} \boldsymbol{\theta}^*_1 \\ \vdots \\ \boldsymbol{\theta}^*_U \end{bmatrix} = \begin{bmatrix} \theta^*_{1,1} & \dots & \theta^*_{1,T} \\ \vdots & \ddots & \vdots \\ \theta^*_{U,1} & \dots & \theta^*_{U,T} \end{bmatrix}$$

A ogni utente u è associata una probabilità, π^T_u, che segue una distribuzione Beta con parametri b^T_1 e b^T_2; essa rappresenta la tendenza dell'utente u a scrivere documenti che trattano di più topic. Queste probabilità sono raccolte in un vettore $U \times 1$ definito come segue

$$\boldsymbol{\pi}^T_{1:U} = \begin{bmatrix} \pi^T_1 \\ \vdots \\ \pi^T_U \end{bmatrix}$$

Gli elementi della matrice $\boldsymbol{\theta}^*_{1:U}$ e del vettore $\boldsymbol{\pi}^T_{1:U}$ sono i parametri delle distribuzioni del topic principale e del tipo dei documenti, queste distribuzioni sono quindi le distribuzioni a priori delle variabili $\mathbf{z}^*$ e $\mathbf{x}$.

3.3.3 Variabili e Distribuzioni a Livello di Documento

Per ogni documento ud è introdotta una variabile indicatrice latente, x_{ud}, che indica se il documento tratta di uno o di più topic:

$$x_{ud} = \begin{cases} 1 & \text{se il documento } ud \text{ tratta di più topic} \\ 0 & \text{se il documento } ud \text{ tratta di un unico topic} \end{cases}$$

La distribuzione condizionata del tipo del documento ud data la tendenza dell'utente u a scrivere documenti che trattano di più topic, π_u^T, segue una distribuzione di Bernoulli con parametro π_u^T:

$$x_{ud}|\pi_u^T \sim Bern(\pi_u^T)$$

Per ogni documento ud è introdotta una variabile latente, z_{ud}^*, che indica il topic principale associato al documento: se il documento ud tratta di un unico topic ($x_{ud} = 0$), le parole non di sottofondo e gli *hashtag* non globali sono generati a partire dal topic principale. $z_{ud}^* \in \{1, \dots, T\}$ è uno scalare che assume il valore t se il topic principale associato al documento ud è il t-mo topic della collezione. La distribuzione condizionata del topic principale del documento ud data la distribuzione sui topic dell'utente u associato al documento, $\boldsymbol{\theta}_u^*$, segue una distribuzione categoriale con vettore di probabilità $\boldsymbol{\theta}_u^*$:

$$z_{ud}^*|\boldsymbol{\theta}_u^* \sim Cat(\boldsymbol{\theta}_u^*)$$

Per ogni documento ud è introdotto un vettore $T \times 1$, detto *vettore dei topic attivi*,

$$\boldsymbol{\lambda}_{ud} = \begin{bmatrix} \lambda_{ud,1} & \dots & \lambda_{ud,t} & \dots & \lambda_{ud,T} \end{bmatrix},$$

dove $\lambda_{ud,t}$ è una variabile indicatrice latente che indica se il topic è attivo nel documento quando quest'ultimo è considerato nella sua rappresentazione come mistura di topic ($x_{ud} = 1$):

$$\lambda_{ud,t} = \begin{cases} 1 & \text{il topic } t \text{ è attivo nel documento } ud \\ 0 & \text{il topic } t \text{ non è attivo nel documento } ud \end{cases}$$

La distribuzione condizionata della presenza del topic t nel documento ud data la probabilità che un topic sia attivo in un documento, δ, segue una distribuzione di Bernoulli con parametro δ:

$$\lambda_{ud,t}|\delta \sim Bern(\delta)$$

L'insieme dei topic attivi nel documento ud è definito come $\boldsymbol{\Lambda}_{ud} = \{t : \lambda_{ud,t} = 1\}$.

A ogni documento d è associata una distribuzione sui T topic, $\boldsymbol{\theta}_{ud}$, che segue una distribuzione di Dirichlet di ordine T con parametro $\boldsymbol{\alpha}^{(ud)}$, ovvero il documento ud è rappresentato da un vettore di probabilità $T \times 1$ in cui

il t-mo elemento indica quanto è importante il t-mo topic della collezione all'interno del documento ud. Il parametro $\boldsymbol{\alpha}^{(ud)}$ dipende dal vettore dei topic attivi del documento ud, $\boldsymbol{\lambda}_{ud}$, ed così definito

$$\boldsymbol{\alpha}^{(ud)} = \alpha_0 \mathbf{1}_T + \boldsymbol{\lambda}_{ud} \circ \boldsymbol{\alpha} = \alpha_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda_{ud,1} \\ \vdots \\ \lambda_{ud,T} \end{bmatrix} \circ \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_T \end{bmatrix} = \begin{bmatrix} \alpha_0 + \lambda_{ud,1}\alpha_1 \\ \vdots \\ \alpha_0 + \lambda_{ud,T}\alpha_T \end{bmatrix}$$

dove $\circ$ indica il *prodotto di Hadamard* e $\mathbf{1}_T$ un vettore $T \times 1$ di soli 1. Questo approccio è un'estensione di quello introdotto in Lin et al., 2014 che, a differenza del metodo originale, può essere utilizzato per definire il parametro di una distribuzione di Dirichlet anche non simmetrica. I parametri α_0 e $\boldsymbol{\alpha}$, detti *weak topic smoothing prior* e *topic smoothing prior*, sono fissati in modo tale che $\alpha_0 \ll \alpha_t$ per $t = 1, \ldots, T$; così facendo, la *weak topic smoothing prior* α_0 definisce la proporzione a priori dei topic non attivi nel *tweet*, mentre il t-mo elemento della *topic smoothing prior*, α_t, definisce la proporzione a priori del topic t quando è attivo poiché, essendo $\alpha_0 \ll \alpha_t$, α_0 risulta trascurabile nella somma $\alpha_0 + \alpha_t$. Inoltre, fissando $\alpha_0 \to 0$,[3] si ha che gli elementi di $\boldsymbol{\theta}_{ud}$ corrispondenti ai topic non attivi, ovvero $\{\theta_{ud,t}\}_{t \notin \boldsymbol{\Lambda}_{ud}}$, assumono valori talmente bassi –ma non nulli– da poter essere assunti pari a zero; sia $\alpha_t^{(ud)}$ il t-mo elemento del vettore $\boldsymbol{\alpha}_t^{(ud)}$, allora si ha che

$$\alpha_t^{(ud)} \approx \begin{cases} \alpha_t & \text{se } t \in \boldsymbol{\Lambda}_{ud} \\ 0 & \text{se } t \notin \boldsymbol{\Lambda}_{ud} \end{cases}$$

Si assume quindi che l'importanza dei topic non attivi all'interno di un documento sia talmente bassa da poter essere considerata nulla; in particolare, la probabilità che parole e *hashtag* abbiano associato un topic non attivo è essenzialmente nulla. Le distribuzioni sui topic dei documenti della collezione sono raccolte in una matrice $D \times T$ definita come segue

$$\boldsymbol{\theta}_{1:D} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_D \end{bmatrix} = \begin{bmatrix} \theta_{1,1} & \ldots & \theta_{1,T} \\ \vdots & \ddots & \vdots \\ \theta_{D,1} & \ldots & \theta_{D,T} \end{bmatrix}$$

dove $D = \sum_{u=1}^{U} D_u$ è il numero totale di documenti nella collezione.

[3] In pratica si fissa $\alpha_0 = 10^{-7}$.

3.3.4 Variabili e Distribuzioni a Livello di Parola

Per ogni parola udn è introdotta una variabile indicatrice latente, y^V_{udn}, che indica se la parola è di sottofondo o è generata a partire da un topic:

$$y^V_{udn} = \begin{cases} 1 & \text{se la parola } udn \text{ è generata a partire da un topic} \\ 0 & \text{se la parola } udn \text{ è di sottofondo} \end{cases}$$

La distribuzione condizionata dell'origine della parola udn data la tendenza generale degli utenti a utilizzare parole appartenenti a una specifica tematica, π^V, segue una distribuzione di Bernoulli con parametro π^V:

$$y^V_{udn}|\pi^V \sim Bern(\pi^V)$$

Per ogni parola udn è introdotta una variabile latente z^V_{udn} che indica il topic associato alla parola. $z^V_{udn} \in \{1, \ldots, T\}$ è uno scalare che assume il valore t se il topic associato alla parola udn è il t-mo topic della collezione. La distribuzione condizionata del topic associato alla parola udn date le proporzioni dei topic del documento a cui appartiene, $\boldsymbol{\theta}_{ud}$, segue una distribuzione categoriale con vettore di probabilità $\boldsymbol{\theta}_{ud}$:

$$z^V_{udn}|\boldsymbol{\theta}_{ud} \sim Cat(\boldsymbol{\theta}_{ud})$$

Infine, la distribuzione condizionata dalla parola udn, w_{udn}, segue una distribuzione categoriale con vettore di probabilità diverso in base all'origine della parola e al tipo del documento in cui è contenuta. Se la parola è generata a partire da un topic ($y^V_{udn} = 1$), allora si distinguono due casi: se il documento tratta di più topic ($x_{ud} = 1$), allora la distribuzione dipende dalla distribuzione sulle parole del topic associato alla parola; se il documento tratta di un unico topic ($x_{ud} = 0$), allora la distribuzione dipende dalla distribuzione sulle parole del topic principale associato al documento. Se la parola è di sottofondo ($y^V_{udn} = 0$), allora la distribuzione dipende dalla distribuzione sulle parole delle parole di sottofondo. Per esprimere in termini probabilistici i tre casi appena introdotti, si assume che la distribuzione condizionata della parola w_{udn} sia una mistura di tre distribuzioni categoriali in cui i pesi sono $y^V_{udn}x_{ud}$, $y^V_{udn}(1 - x_{ud})$ e $(1 - y^V_{udn})$:

$$p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}})$$

$$
\begin{aligned}
=& y^V_{udn} x_{ud} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \\
&+ y^V_{udn}(1-x_{ud}) \prod_{v=1}^{V} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} \\
&+ (1-y^V_{udn}) \prod_{v=1}^{V} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}}
\end{aligned}
$$

Essendo y^V_{udn} e x_{ud} variabili indicatrici, la distribuzione può essere riscritta come segue:

$$
\begin{aligned}
&= \left(\prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \right) \\
&\quad \times \left(\prod_{v=1}^{V} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} \right) \\
&\quad \times \left(\prod_{v=1}^{V} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \right) \\
&= \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \\
&= \begin{cases}
\prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } y^V_{udn}=1 \text{ e } x_{ud}=1 \\
\prod_{v=1}^{V} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } y^V_{udn}=1 \text{ e } x_{ud}=0 \\
\prod_{v=1}^{V} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}} & \text{se } y^V_{udn}=0
\end{cases}
\end{aligned}
$$

Alternativamente, è possibile ottenere esattamente lo stessa distribuzione condizionata considerando la seguente distribuzione per w_{udn}:

$$
w_{udn}|\ldots \sim Cat\left(y^V_{udn} x_{ud} \boldsymbol{\phi}_{z^V_{udn}} + y^V_{udn}(1-x_{ud}) \boldsymbol{\phi}_{z^*_{ud}} + (1-y^V_{udn}) \boldsymbol{\phi}^{\mathcal{B}} \right)
$$

dove $|\ldots$ è un'abbreviazione di $|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}$.

3.3.5 Variabili e Distribuzioni a Livello di Hashtag

Per ogni *hashtag udl* è introdotta una variabile indicatrice latente, y^H_{udl}, che indica se l'*hashtag* è globale o è generato a partire da un topic:

$$y^H_{udl} = \begin{cases} 1 & \text{se l'}\textit{hashtag udl}\text{ è generato a partire da un topic} \\ 0 & \text{se l'}\textit{hashtag udl}\text{ è globale} \end{cases}$$

La distribuzione condizionata dell'origine dell'*hashtag udn* data la tendenza generale degli utenti a utilizzare *hashtag* appartenenti ad una specifica tematica, π^H, segue una distribuzione di Bernoulli con parametro π^H:

$$y^H_{udl}|\pi^H \sim Bern(\pi^H)$$

Per ogni *hashtag udl* è introdotta una variabile latente z^H_{udl} che indica il topic associato all'*hashtag*. $z^H_{udl} \in \{1, \ldots, T\}$ è uno scalare che assume il valore t se il topic associato all'*hashtag udl* è il t-mo topic della collezione. La distribuzione condizionata del topic associato all'*hashtag udl* date le proporzioni dei topic del documento a cui appartiene, $\boldsymbol{\theta}_{ud}$, segue una distribuzione categoriale con vettore di probabilità $\boldsymbol{\theta}_{ud}$:

$$z^H_{udl}|\boldsymbol{\theta}_{ud} \sim Cat(\boldsymbol{\theta}_{ud})$$

Infine, la distribuzione condizionata dall'*hashtag udl*, h_{udl}, segue una distribuzione categoriale con vettore di probabilità diverso in base all'origine dell'*hashtag* ed al tipo del documento in cui è contenuto. Se l'*hashtag* è generato a partire da un topic ($y^H_{udl} = 1$), allora si distinguono due casi: se il documento tratta di più topic ($x_{ud} = 1$), allora la distribuzione dipende dalla distribuzione sugli *hashtag* del topic associato all'*hashtag*; se il documento tratta di un unico topic ($x_{ud} = 0$), allora la distribuzione dipende dalla distribuzione sugli *hashtag* del topic principale associato al documento. Se l'*hashtag* è globale ($y^H_{udl} = 0$), allora la distribuzione dipende dalla distribuzione sugli *hashtag* degli *hashtag globali*. Per esprimere in termini probabilistici i tre casi appena introdotti, si assume che la distribuzione condizionata dell'*hashtag* h_{udl} sia una mistura di tre distribuzioni categoriali in cui i pesi sono $y^H_{udl}x_{ud}$, $y^H_{udl}(1 - x_{ud})$ e $(1 - y^H_{udl})$:

$$p(h_{udl}|y^H_{udl}, z^V_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}})$$

$$= \prod_{h=1}^{H} \psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=1}} \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=0}} (\psi_h^{\mathcal{B}})^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}}$$

$$= \begin{cases} \prod_{h=1}^{H} \psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h}} & \text{se } y^H_{udl}=1 \text{ e } x_{ud}=1 \\ \prod_{h=1}^{H} \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}} & \text{se } y^H_{udl}=1 \text{ e } x_{ud}=0 \\ \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\mathbb{1}_{h_{udl}=h}} & \text{se } y^H_{udl}=0 \end{cases}$$

Alternativamente, è possibile ottenere esattamente lo stessa distribuzione condizionata considerando la seguente distribuzione per h_{udl}:

$$h_{udl}|\ldots \sim Cat\left(y^H_{udl}x_{ud}\boldsymbol{\psi}_{z^H_{udl}} + y^H_{udl}(1-x_{ud})\boldsymbol{\psi}_{z^*_{ud}} + (1-y^V_{udl})\boldsymbol{\psi}^{\mathcal{B}}\right)$$

dove $|\ldots$ è un'abbreviazione di $|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}$.

3.3.6 Notazione Abbreviata

Riprendendo la notazione introdotta per le parole e gli *hashtag*, i tipi e i topic principali dei documenti scritti dell'utente u sono raccolti in

$$\mathbf{x}_u = \{x_{u1}, \ldots, x_{ud}, \ldots, x_{uD_u}\},$$
$$\mathbf{z}^*_u = \{z^*_{u1}, \ldots, z^*_{ud}, \ldots, z^*_{uD_u}\};$$

quelli dell'intera collezione in

$$\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_u, \ldots, \mathbf{x}_U\},$$
$$\mathbf{z}^* = \{\mathbf{z}^*_1, \ldots, \mathbf{z}^*_u, \ldots, \mathbf{z}^*_U\}.$$

Le origini delle parole del documento ud, quelle delle parole dei documenti scritti dall'utente u e quelle delle parole di un qualsiasi documento della collezione sono rispettivamente raccolte in

$$\mathbf{y}^V_{ud} = \{y^V_{ud1}, \ldots, y^V_{udn}, \ldots, y^V_{udN_{ud}}\},$$
$$\mathbf{y}^V_u = \{y^V_{u1}, \ldots, y^V_{ud}, \ldots, y^V_{uD_u}\},$$
$$\mathbf{y}^V = \{\mathbf{y}^V_1, \ldots, \mathbf{y}^V_u, \ldots, \mathbf{y}^V_U\}.$$

Analogamente, i topic associati alle parole sono raccolti in

$$\mathbf{z}^V_{ud} = \{z^V_{ud1}, \ldots, z^V_{udn}, \ldots, z^V_{udN_{ud}}\},$$

$$\mathbf{z}_u^V = \{z_{u1}^V, \ldots, z_{ud}^V, \ldots, z_{uD_u}^V\},$$
$$\mathbf{z}^V = \{\mathbf{z}_1^V, \ldots, \mathbf{z}_u^V, \ldots, \mathbf{z}_U^V\}.$$

La stessa notazione è utilizzata anche per le origini degli *hashtag* e i topic associati ad essi:

$$\mathbf{y}_{ud}^H = \{y_{ud1}^H, \ldots, y_{udl}^H, \ldots, y_{udL_{ud}}^H\},$$
$$\mathbf{y}_u^H = \{y_{u1}^H, \ldots, y_{ud}^H, \ldots, y_{uD_u}^H\},$$
$$\mathbf{y}^H = \{\mathbf{y}_1^H, \ldots, \mathbf{y}_u^H, \ldots, \mathbf{y}_U^H\},$$
$$\mathbf{z}_{ud}^H = \{z_{ud1}^H, \ldots, z_{udl}^H, \ldots, z_{udL_{ud}}^H\},$$
$$\mathbf{z}_u^H = \{z_{u1}^H, \ldots, z_{ud}^H, \ldots, z_{uD_u}^H\},$$
$$\mathbf{z}^H = \{\mathbf{z}_1^H, \ldots, \mathbf{z}_u^H, \ldots, \mathbf{z}_U^H\}.$$

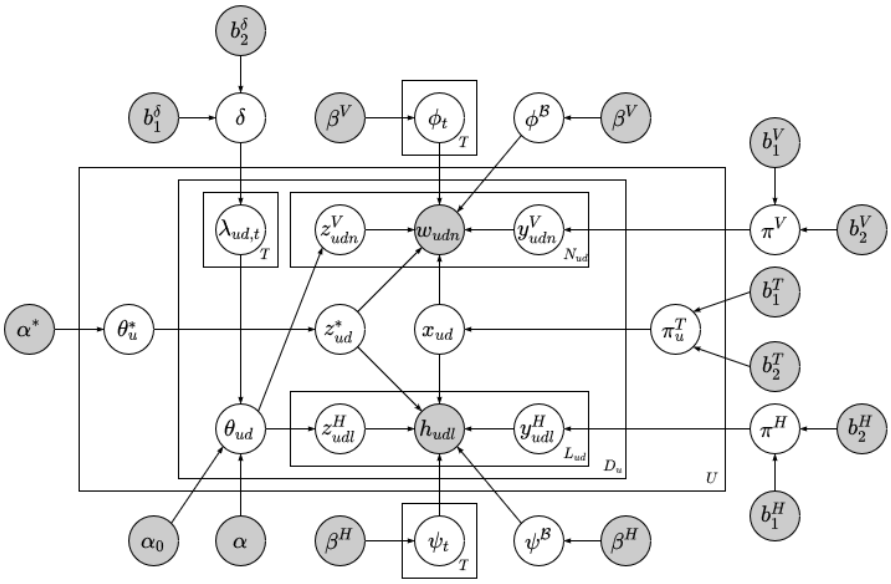


Figura 3.6: *Modello grafico probabilistico* del modello proposto.

3.4 Processo Generativo

Il *processo generativo* del modello proposto, rappresentato in Figura 3.6, è il seguente:

1. Per ogni topic $t = 1, \dots, T$:
   a. Si estrae la distribuzione sulle parole del topic t da una distribuzione di Dirichlet, $\boldsymbol{\phi}_t | \boldsymbol{\beta}^V \sim Dir_V(\boldsymbol{\beta}^V)$.
   b. Si estrae la distribuzione sugli *hashtag* del topic t da una distribuzione di Dirichlet, $\boldsymbol{\psi}_t | \boldsymbol{\beta}^H \sim Dir_H(\boldsymbol{\beta}^H)$.
2. Si estrae la distribuzione sulle parole delle parole di sottofondo da una distribuzione di Dirichlet, $\boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\beta}^V \sim Dir_V(\boldsymbol{\beta}^V)$.
3. Si estrae la distribuzione sugli *hashtag* degli *hashtag globali* da una distribuzione di Dirichlet, $\boldsymbol{\psi}^{\mathcal{B}} | \boldsymbol{\beta}^H \sim Dir_H(\boldsymbol{\beta}^H)$.
4. Si estrae la probabilità che una parola sia generata a partire da un topic da una distribuzione Beta, $\pi^V | b_1^V, b_2^V \sim Beta(b_1^V, b_2^V)$.
5. Si estrae la probabilità che un *hashtag* sia generato a partire da un topic da una distribuzione Beta, $\pi^H | b_1^H, b_2^H \sim Beta(b_1^H, b_2^H)$.
6. Si estrae la probabilità che un topic sia attivo in un documento da una distribuzione Beta, $\delta | b_1^\delta, b_2^\delta \sim Beta(b_1^\delta, b_2^\delta)$.
7. Per ogni utente $u = 1, \dots, U$:
   a. Si estrae la distribuzione sui topic dell'utente u da una distribuzione di Dirichlet, $\boldsymbol{\theta}_u^* | \boldsymbol{\alpha}^* \sim Dir_T(\boldsymbol{\alpha}^*)$.
   b. Si estrae la probabilità che un documento scritto dall'utente u tratti di più topic da una distribuzione Beta, $\pi_u^T | b_1^T, b_2^T \sim Beta(b_1^T, b_2^T)$.
   c. Per ogni *tweet* $d = 1, \dots, D_u$:
      i. Si estrae il tipo del documento ud da una distribuzione di Bernoulli, $x_{ud} | \pi_u^T \sim Bern(\pi_u^T)$.
      ii. Si estrae il topic principale del documento ud, $z_{ud}^* \sim Cat(\boldsymbol{\theta}_u^*)$.
      iii. Per ogni topic t, determino se è attivo nel documento ud, $\lambda_{ud,t} \sim Bern(\delta)$, t=1,...,T.
      iv. Si estrae la distribuzione sui topic del documento ud da una distribuzione di Dirichlet, $\boldsymbol{\theta}_{ud} | \boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha} \sim Dir_T(\boldsymbol{\alpha}^{(ud)})$, dove $\boldsymbol{\alpha}^{(ud)} = \alpha_0 \mathbf{1}_T + \boldsymbol{\lambda}_{ud} \circ \boldsymbol{\alpha}$.

v. Per ogni parola $n = 1, \ldots, N_{ud}$:

A. Si estrae l'origine della parola *udn* da una distribuzione di Bernoulli, $y^V_{udn}|\pi^V \sim Bern(\pi^V)$.

B. Si estrae il topic della parola *udn* dalla distribuzione sui topic del documento *ud*, $z^V_{udn}|\boldsymbol{\theta}_{ud} \sim Cat(\boldsymbol{\theta}_{ud})$.

C. Se $y^V_{udn} = 0$, si estrae la parola *udn* dalla distribuzione sulle parole delle parole di sottofondo,

$$w_{udn}|y^V_{udn} = 0, z^V_{udn}, x_{ud}, z^*_u d, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}} \sim Cat(\boldsymbol{\phi}^{\mathcal{B}});$$

se $y^V_{udn} = 1$ e $x_{ud} = 0$, si estrae la parola *udn* dalla distribuzione sulle parole del topic del documento *ud*,

$$w_{udn}|y^V_{udn} = 1, z^V_{udn}, x_{ud} = 0, z^*_u d, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}} \sim Cat(\boldsymbol{\phi}_{z^*_{ud}});$$

se $y^V_{udn} = 1$ e $x_{ud} = 1$, si estrae la parola *udn* dalla distribuzione sulle parole del topic della parola *udn*,

$$w_{udn}|y^V_{udn} = 1, z^V_{udn}, x_{ud} = 1, z^*_u d, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}} \sim Cat(\boldsymbol{\phi}_{z^V_{udn}}).$$

vi. Per ogni *hashtag* $l = 1, \ldots, L_{ud}$:

A. Si estrae l'origine dell'*hashtag udl* da una distribuzione di Bernoulli, $y^H_{udl}|\pi^H \sim Bern(\pi^H)$.

B. Si estrae il topic dell'*hashtag udl* dalla distribuzione sui topic del documento *ud*, $z^H_{udl}|\boldsymbol{\theta}_{ud} \sim Cat(\boldsymbol{\theta}_{ud})$.

C. Se $y^H_{udl} = 0$, si estrae l'*hashtag udl* dalla distribuzione sugli *hashtag* degli *hashtag globali*,

$$h_{udl}|y^H_{udl} = 0, z^H_{udl}, x_{ud}, z^*_u d, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}} \sim Cat(\boldsymbol{\psi}^{\mathcal{B}});$$

se $y^H_{udl} = 1$ e $x_{ud} = 0$, si estrae l'*hashtag udl* dalla distribuzione sugli *hashtag* del topic del documento *ud*,

$$h_{udl}|y^H_{udl} = 1, z^H_{udl}, x_{ud} = 0, z^*_u d, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}} \sim Cat(\boldsymbol{\psi}_{z^*_{ud}});$$

se $y^V_{udn} = 1$ e $x_{ud} = 1$, si estrae l'*hashtag udl* dalla distribuzione sugli *hashtag* del topic dell'*hashtag udl*,

$$h_{udl}|y^H_{udl} = 1, z^H_{udl}, x_{ud} = 1, z^*_u d, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}} \sim Cat(\boldsymbol{\psi}_{z^H_{udl}}).$$

Variabili osservate		
w_{udn}	$mistura_{w_{udn}}$	n-ma parola del documento ud.
h_{udl}	$mistura_{h_{udl}}$	l-mo *hashtag* del documento ud.
$mistura_{w_{udn}} = y^V_{udn} x_{ud} Cat(\boldsymbol{\phi}_{z^V_{udn}}) + y^V_{udn}(1-x_{ud})Cat(\boldsymbol{\phi}_{z^*_{ud}}) + (1-y^V_{udn})Cat(\boldsymbol{\phi}^{\mathcal{B}})$		
$mistura_{h_{udl}} = y^H_{udl} x_{ud} Cat(\boldsymbol{\psi}_{z^H_{udl}}) + y^H_{udl}(1-x_{ud})Cat(\boldsymbol{\psi}_{z^*_{ud}}) + (1-y^H_{udl})Cat(\boldsymbol{\psi}^{\mathcal{B}})$		

Variabili latenti		
x_{ud}	$Bern(\pi^T_u)$	Indica se il documento ud tratta di più topic.
z^*_{ud}	$Cat(\boldsymbol{\theta}^*_u)$	Topic associato documento ud.
$\lambda_{ud,t}$	$Bern(\delta)$	Indica se il topic t è attivo nel documento ud.
$\boldsymbol{\theta}_{ud}$	$Dir_T(\boldsymbol{\alpha}^{(ud)})$	Distribuzione sui T topic del documento ud.
y^V_{udn}	$Bern(\pi^V)$	Indica se la parola udn ha un topic associato.
z^V_{udn}	$Cat(\boldsymbol{\theta}_{ud})$	Topic associato alla parola udn.
y^H_{udl}	$Bern(\pi^H)$	Indica se l'*hashtag* udl ha un topic associato.
z^H_{udl}	$Cat(\boldsymbol{\theta}_{ud})$	Topic associato all'*hashtag* udl.
$\boldsymbol{\alpha}^{(ud)} = \alpha_0 \mathbf{1}_T + \boldsymbol{\lambda}_{ud}\boldsymbol{\alpha}$		

Parametri		
$\boldsymbol{\phi}_t$	$Dir_V(\boldsymbol{\beta}^V)$	Distribuzione sulle V parole del topic t.
$\boldsymbol{\psi}_t$	$Dir_H(\boldsymbol{\beta}^H)$	Distribuzione sugli H *hashtag* del topic t.
$\boldsymbol{\phi}^{\mathcal{B}}$	$Dir_V(\boldsymbol{\beta}^V)$	Distribuzione sulle V parole delle parole di sottofondo.
$\boldsymbol{\psi}^{\mathcal{B}}$	$Dir_H(\boldsymbol{\beta}^H)$	Distribuzione sugli H *hashtag* degli *hashtag globali*.
π^V	$Beta(b^V_1, b^V_2)$	Probabilità che una parola abbia un topic associato.
π^H	$Beta(b^H_1, b^H_2)$	Probabilità che un *hashtag* abbia un topic associato.
$\boldsymbol{\theta}^*_u$	$Dir_T(\boldsymbol{\alpha}^*)$	Distribuzione sui T topic dell'utente u.
π^T_u	$Beta(b^T_1, b^T_2)$	Probabilità che l'utente u pubblichi un documento con più topic.

Tabella 3.1: Lista delle variabili osservate, delle variabili latenti e dei parametri del modello proposto; le variabili sono ordinate per livello, partendo dalla collezione fino ad arrivare alle parole e agli *hashtag*. L'indice u indica l'utente u, ud il d-mo documento scritto dall'utente u, udn l'n-ma parola del documento ud, udl l'l-mo *hashtag* del documento ud.

3.5 Derivazione della Distribuzione Congiunta

Si espone come ottenere la distribuzione congiunta delle variabili osservate e latenti del modello proposto partendo dalla *Latent Dirichlet Allocation* e poi considerando casi sempre più complessi:

1. il caso di partenza è la *Latent Dirichlet Allocation*;
2. nella prima estensione si introduce una distinzione tra documenti che trattano di un unico topic e di più topic;
3. nelle seconda estensione si introduce la distinzione tra parole generate a partire da un topic e parole di sottofondo;
4. nell'ultima estensione si introducono gli *hashtag* e si associa loro un processo generativo identico a quello delle parole.

Per ogni caso si introducono brevemente le variabili in più rispetto al caso precedente e si espone il procedimento per ottenere la sua distribuzione congiunta. Più nello specifico, prima si introducono le distribuzioni delle variabili latenti che dipendono solo da parametri fissati, poi si ripercorre il processo generativo per determinare la distribuzione congiunta di tutti i documenti della collezione partendo dalla distribuzione di una singola parola, infine si moltiplicano tra loro le distribuzioni ottenute in precedenza per ottenere la distribuzione congiunta delle variabili osservate e latenti del caso considerato.

3.5.1 Caso di Partenza: LDA

Si consideri una generalizzazione della *Latent Dirichlet Allocation* (Blei et al., 2003) in cui le distribuzioni di Dirichlet simmetriche sono sostituite con distribuzioni di Dirichlet. A ogni documento è associata una distribuzione sui T topic, $\boldsymbol{\theta}_{ud}$, che segue una distribuzione di Dirichlet di ordine T con parametro $\boldsymbol{\alpha}$ e ad ogni topic è associata una distribuzione sulle V parole, $\boldsymbol{\phi}_t$, che segue una distribuzione di Dirichlet di parametro $\boldsymbol{\beta}^V$:

$$\begin{aligned} \boldsymbol{\theta}_{ud} &\sim Dir_T(\boldsymbol{\alpha}) \\ \boldsymbol{\phi}_t &\sim Dir_V(\boldsymbol{\beta}^V) \end{aligned}$$

Assumendo che le variabili latenti le cui distribuzioni dipendono solo da parametri noti siano indipendenti tra loro, la distribuzione congiunta delle

variabili a livello di collezione –distribuzioni sulle parole dei topic– e a livello di documento –proporzioni dei topic dei documenti– è data da:

$$\begin{aligned} p(\boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}|\boldsymbol{\alpha}, \boldsymbol{\beta}^V) =& p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\alpha})p(\boldsymbol{\phi}_{1:T}|\boldsymbol{\beta}^V) \\ =& \prod_{u=1}^{U}\prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud}|\boldsymbol{\alpha}) \times \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \\ =& \left(\prod_{u=1}^{U}\prod_{d=1}^{D_u} \frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_t - 1}\right) \\ & \times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V}\beta_v^V)}{\prod_{v=1}^{V}\Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1}\right) \end{aligned}$$

Per ogni parola in ogni documento è introdotta una variabile latente $z^V_{udn} \in \{1, \ldots, T\}$ che segue una distribuzione categoriale il cui vettore di probabilità è la proporzione sui topic del documento in cui essa è contenuta e la parola effettivamente osservata $w_{udn} \in \{1, \ldots, V\}$ segue una distribuzione categoriale il cui vettore di probabilità è la distribuzione sulle parole del topic ad essa associata:

$$\begin{aligned} z^V_{udn}|\boldsymbol{\theta}_{ud} &\sim Cat(\boldsymbol{\theta}_{ud}) \\ w_{udn}|z^V_{udn}, \boldsymbol{\phi}_{1:T} &\sim Cat(\boldsymbol{\phi}_{z^V_{udn}}) \end{aligned}$$

La distribuzione della variabile latente z^V_{udn} dato $\boldsymbol{\theta}_{ud}$ e la distribuzione della variabile osservata w_{udn} dati z^V_{udn} e $\boldsymbol{\phi}_{1:T}$ sono quindi:

$$\begin{aligned} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) &= \prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \\ p(w_{udn}|z^V_{udn}, \boldsymbol{\phi}_{1:T}) &= \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \end{aligned}$$

dove $\mathbb{1}_{prop}$[4] vale 1 se la proposizione *prop* è vera, zero altrimenti:

$$\mathbb{1}_{prop} = \begin{cases} 1 & \text{se } prop \text{ è vera} \\ 0 & \text{altrimenti} \end{cases}$$

La distribuzione congiunta di (z^V_{udn}, w_{udn}) dati $\boldsymbol{\theta}_{ud}$ e $\boldsymbol{\phi}_{1:T}$, ovvero la probabilità di osservare la parola w_{udn} con associato il topic z^V_{udn} date le distribuzioni

[4] Questa è una notazione alternativa della *parentesi di Iverson*, [*prop*].

sulle parole dei T topic $\boldsymbol{\phi}_{1:T}$ e la distribuzione sui topic del documento ud, $\boldsymbol{\theta}_{ud}$, è data da:

$$\begin{aligned} p(z^V_{udn}, w_{udn}|\boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) &= p(z^V_{udn}|\boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) p(w_{udn}|z^V_{udn}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\ &= p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(w_{udn}|z^V_{udn}, \boldsymbol{\phi}_{1:T}) \\ &= \prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \end{aligned}$$

Data l'assunzione di scambiabilità delle parole all'interno di un documento, la distribuzione congiunta condizionata delle variabili relative a un singolo documento è data da:

$$\begin{aligned} p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}|\boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) &= \prod_{n=1}^{N_{ud}} p(z^V_{udn}, w_{udn}|\boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\ &= \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(w_{udn}|z^V_{udn}, \boldsymbol{\phi}_{1:T}) \\ &= \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \right) \end{aligned}$$

Data l'assunzione di scambiabilità dei documenti, la distribuzione congiunta condizionata delle variabili relative ai documenti dell'intera collezione date le loro proporzioni dei topic e le distribuzioni sulle parole dei topic è data da:

$$\begin{aligned} p(\mathbf{z}, \mathbf{w}|\boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}) &= \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}|\boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\ &= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(z^V_{udn}, w_{udn}|\boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\ &= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(w_{udn}|z^V_{udn}, \boldsymbol{\phi}_{1:T}) \\ &= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \right) \end{aligned}$$

Infine, la distribuzione congiunta delle variabili osservate e latenti dati i parametri fissati $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}^V$ è data da:

$$p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}|\boldsymbol{\alpha}, \boldsymbol{\beta}^V)$$

$$
\begin{aligned}
&= p(\mathbf{z}^V, \mathbf{w} | \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\alpha}, \boldsymbol{\beta}^V) p(\boldsymbol{\theta}_{1:U}, \boldsymbol{\phi}_{1:T} | \boldsymbol{\alpha}, \boldsymbol{\beta}^V) \\
&= p(\mathbf{z}^V, \mathbf{w} | \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\alpha}, \boldsymbol{\beta}^V) p(\boldsymbol{\theta}_{1:D} | \boldsymbol{\alpha}, \boldsymbol{\beta}^V) p(\boldsymbol{\phi}_{1:T} | \boldsymbol{\alpha}, \boldsymbol{\beta}^V) \\
&= p(\mathbf{z}^V, \mathbf{w} | \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}) p(\boldsymbol{\theta}_{1:D} | \boldsymbol{\alpha}) p(\boldsymbol{\phi}_{1:T} | \boldsymbol{\beta}^V) \\
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(z^V_{udn} | \boldsymbol{\theta}_{ud}) p(w_{udn} | z^V_{udn}, \boldsymbol{\phi}_{1:T}) \\
&\quad \times \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud} | \boldsymbol{\alpha}) \times \prod_{t=1}^{T} p(\boldsymbol{\phi}_t | \boldsymbol{\beta}^V) \\
&= \prod_{t=1}^{T} p(\boldsymbol{\phi}_t | \boldsymbol{\beta}^V) \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud} | \boldsymbol{\alpha}) \prod_{n=1}^{N_{ud}} p(z^V_{udn} | \boldsymbol{\theta}_{ud}) p(w_{udn} | z^V_{udn}, \boldsymbol{\phi}_{1:T}) \\
&= \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \left(\frac{\Gamma(\sum_{t=1}^{T} \alpha_t)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_t - 1} \right) \\
&\quad \times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \right)
\end{aligned}
$$

3.5.2 Documenti che Trattano di un unico Topic e di più Topic

Si consideri un'estensione del caso precedente in cui ogni documento può trattare di un unico topic con probabilità $1-\pi_u^T$ o di più topic con probabilità π_u^T, dove π_u^T rappresenta la tendenza dell'utente u a scrivere documenti che trattano di più topic:

1. nel primo caso si associa un topic, detto topic principale, all'intero documento e tutte le parole seguono una distribuzione categoriale il cui vettore di probabilità è la distribuzione sulle parole del topic principale;
2. nel secondo caso si seleziona un sottogruppo dei topic e si genera la distribuzione sui topic del documento, il topic associato a ogni parola segue una distribuzione categoriale il cui vettore di probabilità è la distribuzione sui topic del documento e infine ogni parola segue una distribuzione categoriale il cui vettore di probabilità è la distribuzione sulle parole del topic associato ad essa.

Si assume quindi che ogni utente e ogni documento abbiano associati una distribuzione sui T topic: quelle degli utenti sono utilizzate per determinare il topic principale di ogni documento, mentre quelle dei documenti sono

utilizzate per determinare i topic assegnati alle parole. In particolare, a ogni utente u è associata:

- la distribuzione sui T topic $\boldsymbol{\theta}_u^*$, che rappresenta le preferenze in termini di topic dell'utente quando decide di scrivere un documento con un unico topic; $\boldsymbol{\theta}_u^*$ segue una distribuzione di Dirichlet di parametro $\boldsymbol{\alpha}^*$,

$$\boldsymbol{\theta}_u^* \sim Dir_T(\boldsymbol{\alpha^*}).$$

Assumendo che tutte le variabili latenti le cui distribuzioni dipendono solo da parametri noti siano indipendenti tra loro, la distribuzione congiunta delle variabili a livello di collezione –le distribuzioni sulle parole dei topic e la probabilità che un topic sia attivo in un documento– e a livello di utente –le distribuzioni sui topic e le tendenze a scrivere documenti con più topic degli utenti– è data da:

$$\begin{aligned}
&p(\boldsymbol{\theta}_{1:U}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}_{1:U}^T, \delta | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^V, b_1^T, b_2^T, b_1^\delta, b_2^\delta) \\
&= p(\boldsymbol{\theta}_{1:U}^* | \boldsymbol{\alpha}^*) p(\boldsymbol{\phi}_{1:T} | \boldsymbol{\beta}^V) p(\boldsymbol{\pi}_{1:U}^T | b_1^T, b_2^T) p(\delta | b_1^\delta, b_2^\delta) \\
&= \left(\prod_{u=1}^{U} p(\boldsymbol{\theta}_u^* | \boldsymbol{\alpha}^*) \right) \left(\prod_{t=1}^{T} p(\boldsymbol{\phi}_t | \boldsymbol{\beta}^V) \right) \left(\prod_{u=1}^{U} p(\pi_u^T | b_1^T, b_2^T) \right) p(\delta | b_1^\delta, b_2^\delta) \\
&= \left(\prod_{u=1}^{U} \frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \right) \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\quad \times \left(\prod_{u=1}^{U} \frac{\Gamma(b_1^T + b_2^T)}{\Gamma(b_1^T)\Gamma(b_2^T)} (\pi_u^T)^{b_1^T - 1} (1 - \pi_u^T)^{b_2^T - 1} \right) \\
&\quad \times \left(\frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)} \delta^{b_1^\delta - 1} (1 - \delta)^{b_2^\delta - 1} \right)
\end{aligned}$$

A ogni documento ud sono associate quattro variabili latenti:

- il tipo del documento $x_{ud} \in \{0, 1\}$, che indica se il documento tratta di uno ($x_{ud} = 0$) o di più topic ($x_{ud} = 1$), segue una distribuzione di Bernoulli di parametro π_u^T; la distribuzione a priori della probabilità è $\pi_u^T \sim Beta(b_1^T, b_2^T)$;
- il topic principale $z_{ud}^* \in \{1, \ldots, T\}$, che viene utilizzato per generare una parola se il documento tratta di un unico topic ($x_{ud} = 0$), segue una distribuzione categoriale il cui vettore di probabilità è la distribuzione sui topic dell'utente $\boldsymbol{\theta}_u^*$;

- il vettore dei topic attivi $\boldsymbol{\lambda}_{ud} = (\lambda_{ud,1} \dots \lambda_{ud,T})$ indica di quali topic il documento tratta se il documento tratta di più topic ($x_{ud} = 1$); ogni variabile $\lambda_{ud,t} \in \{0, 1\}$ segue una distribuzione di Bernoulli di parametro δ; la distribuzione a priori della probabilità è $\delta \sim Beta(b_1^\delta, b_2^\delta)$;
- la distribuzione sui T topic $\boldsymbol{\theta}_{ud}$, che rappresenta la proporzione dei topic del documento, segue una distribuzione di Dirichlet di parametro $\boldsymbol{\alpha}^{(ud)} = \alpha_0 \mathbf{1}_T + \boldsymbol{\lambda}_{ud} \circ \boldsymbol{\alpha}$, $\boldsymbol{\theta}_{ud} \sim Dir_T(\boldsymbol{\alpha}^{(ud)})$.

Le distribuzioni condizionate delle variabili latenti x_{ud}, z^*_{ud}, $\boldsymbol{\lambda}_{ud}$ e $\boldsymbol{\theta}_{ud}$ sono:

$$
\begin{aligned}
p(x_{ud}|\pi_u^T) &= (\pi_u^T)^{x_{ud}}(1-\pi_u^T)^{1-x_{ud}} \\
p(z^*_{ud}|\boldsymbol{\theta}^*_u) &= \prod_{t=1}^{T}(\theta^*_{u,t})^{\mathbb{1}_{z^*_{ud}=t}} \\
p(\boldsymbol{\lambda}_{ud}|\delta) &= \prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) \\
&= \prod_{t=1}^{T} \delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}} \\
p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) &= \frac{\Gamma(\sum_{t=1}^{T}\alpha_t^{(ud)})}{\prod_{t=1}^{T}\Gamma(\alpha_t^{(ud)})}\prod_{t=1}^{T}\theta_{ud,t}^{\alpha_t^{(ud)}-1} \\
&= \frac{\Gamma(\sum_{t=1}^{T}(\alpha_0+\lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t)}\prod_{t=1}^{T}\theta_{ud,t}^{\alpha_0+\lambda_{ud,t}\alpha_t-1}
\end{aligned}
$$

Per esprimere in termini probabilistici i due casi introdotti in precedenza, si assume che la distribuzione di una parola w_{udn} dati z^V_{udn}, $\boldsymbol{\phi}_{1:T}$, x_{ud} e z^*_{ud} sia una mistura di due distribuzioni categoriali in cui i pesi sono x_{ud} e $1 - x_{ud}$; essendo x_{ud} una variabile indicatrice, si considera sempre o la prima o la seconda distribuzione:

$$
\begin{aligned}
p(w_{udn}|z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}) =& x_{ud}\prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} + (1-x_{ud})\prod_{v=1}^{V}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} \\
&= \prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=1}}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=0}} \\
&= \begin{cases} \prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } x_{ud}=1 \\ \prod_{v=1}^{V}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } x_{ud}=0 \end{cases}
\end{aligned}
$$

Seguendo l'intuizione di Griffiths et al., 2005, a tutti i documenti sono associate tutte le variabili latenti in precedenza introdotte –x_{ud}, z^*_{ud}, $\boldsymbol{\lambda}_{ud}$ e $\boldsymbol{\theta}_{ud}$– e a tutte le parole è associato un topic z^V_{udn}: solo la distribuzione di una parola dipende esclusivamente dalle variabili legate al processo generativo del tipo del documento in cui è contenuta. Ciò è evidente osservando la distribuzione congiunta di (z^V_{udn}, w_{udn}) dati x_{ud}, z^*_{ud}, $\boldsymbol{\theta}_{ud}$ e $\boldsymbol{\phi}_{1:T}$, che dipende dal topic z^V_{udn} assegnato alla parola anche quando quest'ultima non è effettivamente influenzata da esso ($x_{ud} = 0$):

$$
\begin{aligned}
&p(z^V_{udn}, w_{udn} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\
=&p(z^V_{udn} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) p(w_{udn} | z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\
=&p(z^V_{udn} | \boldsymbol{\theta}_{ud}) p(w_{udn} | z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}) \\
=&\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=0}} \\
=&\begin{cases}
\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } x_{ud} = 1 \\
\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}=t}} \prod_{v=1}^{V} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } x_{ud} = 0
\end{cases}
\end{aligned}
$$

Data l'assunzione di scambiabilità delle parole all'interno di un documento, la distribuzione congiunta condizionata delle variabili relative a un singolo documento è data da:

$$
\begin{aligned}
&p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, x_{ud}, z^*_{ud}, \boldsymbol{\lambda}_{ud}, \boldsymbol{\theta}_{ud} | \boldsymbol{\theta}^*_u, \boldsymbol{\phi}_{1:T}, \pi^T_u, \delta, \boldsymbol{\alpha}, \alpha_0) \\
=&p(x_{ud} | \pi^T_u) p(z^*_{ud} | \boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud} | \delta) p(\boldsymbol{\theta}_{ud} | \boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\
=&p(x_{ud} | \pi^T_u) p(z^*_{ud} | \boldsymbol{\theta}^*_u) \prod_{t=1}^{T} p(\lambda_{ud,t} | \delta) p(\boldsymbol{\theta}_{ud} | \boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}, w_{udn} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T})
\end{aligned}
$$

Data l'assunzione di scambiabilità dei documenti scritti da uno stesso utente, la distribuzione congiunta condizionata dell'intera collezione è data da:

$$
p(\mathbf{z}^V, \mathbf{w}, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D} | \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \boldsymbol{\alpha}, \alpha_0)
$$

$$
\begin{aligned}
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\mathbf{z}_{ud}^V, \mathbf{w}_{ud}, x_{ud}, z_{ud}^*, \boldsymbol{\lambda}_{ud}, \boldsymbol{\theta}_{ud}^* | \boldsymbol{\theta}_u^*, \boldsymbol{\phi}_{1:T}, \pi_u^T, \delta, \boldsymbol{\alpha}, \alpha_0) \\
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T) p(z_{ud}^*|\boldsymbol{\theta}_u^*) \prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\quad \times \prod_{n=1}^{N_{ud}} p(z_{udn}^V, w_{udn}|x_{ud}, z_{ud}^*, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \left((\pi_u^T)^{x_{ud}} (1-\pi_u^T)^{1-x_{ud}} \right) \left(\prod_{t=1}^{T} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^*=t}} \right) \left(\prod_{t=1}^{T} \delta^{\lambda_{ud,t}} (1-\delta)^{1-\lambda_{ud,t}} \right) \\
&\quad \times \left(\frac{\Gamma(\sum_{t=1}^{T} (\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1} \right) \\
&\quad \times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z_{udn}^V=t}} \prod_{v=1}^{V} \phi_{z_{udn}^V,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=1}} \phi_{z_{ud}^*,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=0}} \right)
\end{aligned}
$$

Infine, la distribuzione congiunta delle variabili osservate e latenti dati i parametri fissati è data da:

$$
\begin{aligned}
&p(\mathbf{z}^V, \mathbf{w}, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\theta}_{1:U}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}_{1:U}^T, \delta | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, b_1^T, b_2^T, b_1^\delta, b_2^\delta) \\
&= p(\mathbf{z}^V, \mathbf{w}, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D} | \boldsymbol{\theta}_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}_{1:U}^T, \delta, \boldsymbol{\alpha}, \alpha_0) \\
&\quad \times p(\boldsymbol{\theta}_{1:U}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}_{1:U}^T, \delta | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^V, b_1^T, b_2^T, b_1^\delta, b_2^\delta) \\
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T) p(z_{ud}^*|\boldsymbol{\theta}_u^*) \prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\quad \times \prod_{n=1}^{N_{ud}} p(z_{udn}^V, w_{udn}|x_{ud}, z_{ud}^*, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}) \\
&\quad \times \left(\prod_{u=1}^{U} p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*) \right) \left(\prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \right) \left(\prod_{u=1}^{U} p(\pi_u^T|b_1^T, b_2^T) \right) p(\delta|b_1^\delta, b_2^\delta) \\
&= p(\delta|b_1^\delta, b_2^\delta) \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \prod_{u=1}^{U} p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*) p(\pi_u^T|b_1^T, b_2^T) \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T) p(z_{ud}^*|\boldsymbol{\theta}_u^*) \\
&\quad \times p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z_{udn}^V|\boldsymbol{\theta}_{ud}) p(w_{udn}|z_{udn}^V, x_{ud}, z_{ud}^*, \boldsymbol{\phi}_{1:T}) \\
&= \left(\frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)} \delta^{b_1^\delta - 1} (1-\delta)^{b_2^\delta - 1} \right) \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \boldsymbol{\beta}_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\quad \times \prod_{u=1}^{U} \left(\frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \right) \left(\frac{\Gamma(b_1^T + b_2^T)}{\Gamma(b_1^T)\Gamma(b_2^T)} (\pi_u^T)^{b_1^T - 1} (1-\pi_u^T)^{b_2^T - 1} \right)
\end{aligned}
$$

$$\times \prod_{d=1}^{D_u} \left((\pi_u^T)^{x_{ud}} (1-\pi_u^T)^{1-x_{ud}} \right) \left(\prod_{t=1}^{T} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^*=t}} \right)$$
$$\times \left(\prod_{t=1}^{T} \delta^{\lambda_{ud,t}} (1-\delta)^{1-\lambda_{ud,t}} \right) \left(\frac{\Gamma(\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1} \right)$$
$$\times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z_{udn}^V=t}} \prod_{v=1}^{V} \phi_{z_{udn}^V,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=1}} \phi_{z_{ud}^*,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{x_{ud}=0}} \right)$$

3.5.3 Parole di Sottofondo

Si consideri un'estensione del caso precedente in cui ogni parola può essere generata a partire da un topic con probabilità π^V oppure essere di sottofondo con probabilità $1-\pi^V$, dove π^V indica la probabilità che una parola sia generata a partire da un topic. Nel primo caso una parola è generata come nella sezione precedente, attraverso una mistura di distribuzioni; nel secondo caso una parola è estratta da una distribuzione categoriale il cui vettore di probabilità è la distribuzione sulle parole delle parole di sottofondo. Quest'ultima, $\boldsymbol{\phi}^{\mathcal{B}}$, segue una distribuzione di Dirichlet di parametro $\boldsymbol{\beta}^V$, $\boldsymbol{\phi}^{\mathcal{B}} \sim Dir_V(\boldsymbol{\beta}^V)$, da cui

$$\begin{aligned}
&p(\boldsymbol{\theta}_{1:U}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}_{1:U}^T, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^V, b_1^T, b_2^T, b_1^\delta, b_2^\delta, b_1^V, b_2^V) \\
=&p(\boldsymbol{\theta}_{1:U}^* | \boldsymbol{\alpha}^*) p(\boldsymbol{\phi}_{1:T} | \boldsymbol{\beta}^V) p(\boldsymbol{\pi}_{1:U}^T | b_1^T, b_2^T) p(\delta | b_1^\delta, b_2^\delta) p(\pi^V | b_1^V, b_2^V) p(\boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\beta}^V) \\
=&p(\delta | b_1^\delta, b_2^\delta) p(\pi^V | b_1^V, b_2^V) p(\boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\beta}^V) \prod_{t=1}^{T} p(\boldsymbol{\phi}_t | \boldsymbol{\beta}^V) \prod_{u=1}^{U} p(\boldsymbol{\theta}_u^* | \boldsymbol{\alpha}^*) p(\pi_u^T | b_1^T, b_2^T)
\end{aligned}$$

dove le due distribuzioni non ancora introdotte sono

$$p(\pi^V | b_1^V, b_2^V) = \frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)} (\pi^V)^{b_1^V - 1} (1-\pi^V)^{b_2^V - 1}$$

e

$$p(\boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\beta}^V) = \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} (\phi_v^{\mathcal{B}})^{\beta_v^V - 1}.$$

In aggiunta alle variabili aleatorie introdotte nel caso precedente, a ogni parola è associata:

- l'origine della parola $y^V_{udn} \in \{0,1\}$, che indica se la parola è generata a partire da un topic ($y^V_{udn} = 1$) o se è di sottofondo ($y^V_{udn} = 0$), segue una distribuzione di Bernoulli di parametro π^V; la distribuzione a priori della probabilità è $\pi^V \sim Beta(b^V_1, b^V_2)$.

La distribuzione della variabile latente y^V_{udn} data la probabilità π^V è:

$$p(y^V_{udn}|\pi^V) = (\pi^V)^{y^V_{udn}}(1-\pi^V)^{1-y^V_{udn}}$$

Per esprimere in termini probabilistici i due casi introdotti in precedenza, si assume che la distribuzione di una parola w_{udn} sia una mistura di due distribuzioni in cui i pesi sono y^V_{udn} e $(1-y^V_{udn})$; essendo y^V_{udn} una variabile indicatrice, si considera sempre o la prima o la seconda distribuzione:

$$\begin{aligned}
&p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
=&y^V_{udn}\left[x_{ud}\prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} + (1-x_{ud})\prod_{v=1}^{V}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}}\right] \\
&+(1-y^V_{udn})\prod_{v=1}^{V}(\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}}
\end{aligned}$$

Essendo la prima distribuzione della mistura con pesi y^V_{udn} e $1-y^V_{udn}$ a sua volta una mistura con pesi x_{ud} e $1-x_{ud}$, si possono distinguere tre casi:

$$\begin{aligned}
=&y^V_{udn}x_{ud}\prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \\
&+y^V_{udn}(1-x_{ud})\prod_{v=1}^{V}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} \\
&+(1-y^V_{udn})\prod_{v=1}^{V}(\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}} \\
=&\prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}}\left(\phi^{\mathcal{B}}_v\right)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}}
\end{aligned}$$

$$= \begin{cases} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } y^V_{udn}=1 \text{ e } x_{ud}=1 \\ \prod_{v=1}^{V} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} & \text{se } y^V_{udn}=1 \text{ e } x_{ud}=0 \\ \prod_{v=1}^{V} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}} & \text{se } y^V_{udn}=0 \end{cases}$$

La distribuzione congiunta condizionata delle tre variabili aleatorie a livello di parola, $(z^V_{udn}, w_{udn}, y^V_{udn})$, è data da:

$$\begin{aligned} & p(z^V_{udn}, w_{udn}, y^V_{udn}|x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\ =& p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(w_{udn}, y^V_{udn}|z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\ =& p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\ =& \prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}}} (\pi^V)^{y^V_{udn}} (1-\pi^V)^{1-y^V_{udn}} \\ & \times \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} \left(\phi^{\mathcal{B}}_v\right)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \end{aligned}$$

Data l'assunzione di scambiabilità delle parole all'interno di un documento, la distribuzione congiunta condizionata delle variabili relative a un singolo documento è data da:

$$\begin{aligned} & p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}, x_{ud}, z^*_{ud}, \boldsymbol{\lambda}_{ud}, \boldsymbol{\theta}_{ud}|\boldsymbol{\theta}^*_u, \boldsymbol{\phi}_{1:T}, \pi^T_u, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\alpha}, \alpha_0) \\ =& p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\ & \times p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}|x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\ =& p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\ & \times \prod_{n=1}^{N_{ud}} p(z^V_{udn}, w_{udn}, y^V_{udn}|x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\ =& p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\ & \times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \end{aligned}$$

Data l'assunzione di scambiabilità dei documenti scritti da uno stesso utente, la distribuzione congiunta condizionata dell'intera collezione è data da:

$$p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D}|\boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\alpha}, \alpha_0)$$

$$
\begin{aligned}
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}, x_{ud}, z^*_{ud}, \boldsymbol{\lambda}_{ud}, \boldsymbol{\theta}_{ud} | \boldsymbol{\theta}^*_u, \boldsymbol{\phi}_{1:T}, \pi^T_u, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\alpha}, \alpha_0) \\
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\quad \times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \left((\pi^T_u)^{x_{ud}} (1-\pi^T_u)^{1-x_{ud}} \right) \left(\prod_{t=1}^{T} (\theta^*_{u,t})^{\mathbb{1}_{z^*_{ud}=t}} \right) \\
&\quad \times \left(\prod_{t=1}^{T} \delta^{\lambda_{ud,t}} (1-\delta)^{1-\lambda_{ud,t}} \right) \left(\frac{\Gamma(\sum_{t=1}^{T} (\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1}_{ud,t} \right) \\
&\quad \times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta^{\mathbb{1}_{z^V_{udn}}}_{ud,t} (\pi^V)^{y^V_{udn}} (1-\pi^V)^{1-y^V_{udn}} \right. \\
&\quad \left. \times \prod_{v=1}^{V} \phi^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{x_{ud}=1}}_{z^V_{udn},v} \phi^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{x_{ud}=0}}_{z^*_{ud},v} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=0}} \right)
\end{aligned}
$$

Infine, la distribuzione congiunta delle variabili osservate e latenti dati i parametri fissati è data da:

$$
\begin{aligned}
&p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}^*_{1:D}, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \mathbf{b}) \\
=&p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D} | \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\alpha}) \\
&\times p(\boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^V, \mathbf{b}) \\
=&\prod_{u=1}^{U} \prod_{d=1}^{D_u} p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(\delta|b^\delta_1, b^\delta_2) p(\pi^V|b^V_1, b^V_2) p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V) p(\boldsymbol{\phi}_{1:T}|\boldsymbol{\beta}^V) p(\boldsymbol{\theta}^*_{1:U}|\boldsymbol{\alpha}^*) p(\boldsymbol{\pi}^T_{1:U}|b^T_1, b^T_2) \\
=&p(\delta|b^\delta_1, b^\delta_2) p(\pi^V|b^V_1, b^V_2) p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V) \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \prod_{u=1}^{U} p(\boldsymbol{\theta}^*_u|\boldsymbol{\alpha}) p(\pi^T_u|b^T_1, b^T_2) \\
&\times \prod_{d=1}^{D_u} p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}})
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)}\delta^{b_1^\delta - 1}(1-\delta)^{b_2^\delta - 1}\right)\left(\frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)}(\pi^V)^{b_1^V-1}(1-\pi^V)^{b_2^V-1}\right)\\
&\quad\times\left(\frac{\Gamma(\sum_{v=1}^V \beta_v^V)}{\prod_{v=1}^V\Gamma(\beta_v^V)}\prod_{v=1}^V(\phi_v^{\mathcal{B}})^{\beta_v^V-1}\right)\left(\prod_{t=1}^T\frac{\Gamma(\sum_{v=1}^V \boldsymbol{\beta}_v^V)}{\prod_{v=1}^V\Gamma(\beta_v^V)}\prod_{v=1}^V\phi_{t,v}^{\beta_v^V-1}\right)\\
&\quad\times\prod_{u=1}^U\left(\frac{\Gamma(\sum_{t=1}^T\alpha_t^*)}{\prod_{t=1}^T\Gamma(\alpha_t^*)}\prod_{t=1}^T(\theta_{u,t}^*)^{\alpha_t^*-1}\right)\left(\frac{\Gamma(b_1^T+b_2^T)}{\Gamma(b_1^T)\Gamma(b_2^T)}(\pi_u^T)^{b_1^T-1}(1-\pi_u^T)^{b_2^T-1}\right)\\
&\quad\times\prod_{d=1}^{D_u}\left((\pi_u^T)^{x_{ud}}(1-\pi_u^T)^{1-x_{ud}}\right)\left(\prod_{t=1}^T(\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^*=t}}\right)\\
&\quad\times\left(\prod_{t=1}^T\delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}}\right)\left(\frac{\Gamma(\sum_{t=1}^T(\alpha_0+\lambda_{ud,t}\alpha_t))}{\prod_{t=1}^T\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t)}\prod_{t=1}^T\theta_{ud,t}^{\alpha_0+\lambda_{ud,t}\alpha_t-1}\right)\\
&\quad\times\prod_{n=1}^{N_{ud}}\left(\prod_{t=1}^T\theta_{ud,t}^{\mathbb{1}_{z_{udn}^V}}(\pi^V)^{y_{udn}^V}(1-\pi^V)^{1-y_{udn}^V}\right.\\
&\quad\left.\times\prod_{v=1}^V\phi_{z_{udn}^V,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y_{udn}^V=1}\mathbb{1}_{x_{ud}=1}}\phi_{z_{ud}^*,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y_{udn}^V=1}\mathbb{1}_{x_{ud}=0}}(\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y_{udn}^V=0}}\right)
\end{aligned}
$$

dove $\mathbf{b} = (b_1^T, b_2^T, b_1^\delta, b_2^\delta, b_1^V, b_2^V)$.

3.5.4 Modello Finale: Parole e Hashtag

Infine, per ottenere la distribuzione del modello proposto in questa tesi, si introducono gli *hashtag* e si associa loro lo stesso processo generativo delle parole: a livello pratico si considerano le stesse distribuzioni di probabilità con parametri diversi per le quantità corrispondenti. Essendo il modello finale, si riportano nuovamente anche le distribuzioni introdotte nei casi più semplici.

Assumendo che tutte le variabili latenti le cui distribuzioni dipendono solo da parametri noti siano indipendenti tra loro, la distribuzione congiunta delle variabili a livello di collezione e a livello di utente è data da:

$$
\begin{aligned}
&p(\boldsymbol{\theta}_{1:U}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}_{1:U}^T, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}} | \boldsymbol{\alpha}^*, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})\\
=&p(\boldsymbol{\theta}_{1:U}^*|\boldsymbol{\alpha}^*)p(\boldsymbol{\phi}_{1:T}|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}_{1:T}|\boldsymbol{\beta}^H)p(\boldsymbol{\pi}_{1:U}^T|b_1^T, b_2^T)\\
&\times p(\delta|b_1^\delta, b_2^\delta)p(\pi^V|b_1^V, b_2^V)p(\pi^H|b_1^H, b_2^H)p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H)\\
=&p(\delta|b_1^\delta, b_2^\delta)p(\pi^V|b_1^V, b_2^V)p(\pi^H|b_1^H, b_2^H)p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H)\\
&\times\prod_{t=1}^T p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}_t|\boldsymbol{\beta}^H)\prod_{u=1}^U p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*)p(\pi_u^T|b_1^T, b_2^T)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)}\delta^{b_1^\delta - 1}(1-\delta)^{b_2^\delta - 1}\right) \\
&\times \left(\frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)}(\pi^V)^{b_1^V - 1}(1-\pi^V)^{b_2^V - 1}\right) \\
&\times \left(\frac{\Gamma(b_1^H + b_2^H)}{\Gamma(b_1^H)\Gamma(b_2^H)}(\pi^H)^{b_1^H - 1}(1-\pi^H)^{b_2^H - 1}\right) \\
&\times \left(\prod_{t=1}^{T}\frac{\Gamma(\sum_{v=1}^{V}\beta_v^V)}{\prod_{v=1}^{V}\Gamma(\beta_v^V)}\prod_{v=1}^{V}\phi_{t,v}^{\beta_v^V - 1}\right)\left(\prod_{t=1}^{T}\frac{\Gamma(\sum_{h=1}^{H}\beta_h^H)}{\prod_{h=1}^{V}\Gamma(\beta_h^H)}\prod_{h=1}^{H}\psi_{t,h}^{\beta_h^H - 1}\right) \\
&\times \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v^V)}{\prod_{v=1}^{V}\Gamma(\beta_v^V)}\prod_{v=1}^{V}(\phi_v^{\mathcal{B}})^{\beta_v^V - 1}\right)\left(\frac{\Gamma(\sum_{h=1}^{H}\beta_h^H)}{\prod_{h=1}^{V}\Gamma(\beta_h^H)}\prod_{h=1}^{H}(\psi_h^{\mathcal{B}})^{\beta_h^H - 1}\right) \\
&\times \prod_{u=1}^{U}\left(\frac{\Gamma(\sum_{t=1}^{T}\alpha_t^*)}{\prod_{t=1}^{T}\Gamma(\alpha_t^*)}\prod_{t=1}^{T}(\theta_{u,t}^*)^{\alpha_t^* - 1}\right)\left(\frac{\Gamma(b_1^T + b_2^T)}{\Gamma(b_1^T)\Gamma(b_2^T)}(\pi_u^T)^{b_1^T - 1}(1-\pi_u^T)^{b_2^T - 1}\right)
\end{aligned}
$$

dove $\mathbf{b} = (b_1^T, b_2^T, b_1^\delta, b_2^\delta, b_1^V, b_2^V, b_1^H, b_2^H)$. Le distribuzioni condizionate delle variabili latenti a livello di documento, x_{ud}, z_{ud}^*, $\boldsymbol{\lambda}_{ud}$, sono:

$$
\begin{aligned}
p(x_{ud}|\pi_u^T) &= (\pi_u^T)^{x_{ud}}(1-\pi_u^T)^{1-x_{ud}} \\
p(z_{ud}^*|\boldsymbol{\theta}_u^*) &= \prod_{t=1}^{T}(\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^*=t}} \\
p(\boldsymbol{\lambda}_{ud}|\delta) &= \prod_{t=1}^{T}p(\lambda_{ud,t}|\delta) = \prod_{t=1}^{T}\delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}} \\
p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) &= \frac{\Gamma(\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)}\prod_{t=1}^{T}\theta_{ud,t}^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1}
\end{aligned}
$$

quelle delle variabili a livello di parola, y_{udn}^V e z_{udn}^V, sono:

$$
\begin{aligned}
p(y_{udn}^V|\pi^V) &= (\pi^V)^{y_{udn}^V}(1-\pi^V)^{1-y_{udn}^V} \\
p(z_{udn}^V|\boldsymbol{\theta}_{ud}) &= \prod_{t=1}^{T}\theta_{ud,t}^{\mathbb{1}_{z_{udn}^V=t}}
\end{aligned}
$$

quelle delle variabili a livello di *hashtag*, y_{udl}^H e z_{udl}^H, sono:

$$
\begin{aligned}
p(y_{udl}^H|\pi^H) &= (\pi^H)^{y_{udl}^H}(1-\pi^H)^{1-y_{udl}^H} \\
p(z_{udl}^H|\boldsymbol{\theta}_{ud}) &= \prod_{t=1}^{T}\theta_{ud,t}^{\mathbb{1}_{z_{udl}^H=t}}
\end{aligned}
$$

Le distribuzioni condizionate delle variabili osservate w_{udn} e h_{udl} sono:

$$
\begin{aligned}
&p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&= \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \\
&p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) \\
&= \prod_{h=1}^{H} \psi_{z^H_{udl},v}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=1}} \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=0}} (\psi^{\mathcal{B}}_h)^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}}
\end{aligned}
$$

La distribuzione congiunta condizionata delle tre variabili aleatorie a livello di parola, $(z^V_{udn}, w_{udn}, y^V_{udn})$, è data da:

$$
\begin{aligned}
&p(z^V_{udn}, w_{udn}, y^V_{udn}|x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\
=&p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(w_{udn}, y^V_{udn}|z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\
=&p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
=&\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}}} (\pi^V)^{y^V_{udn}} (1-\pi^V)^{1-y^V_{udn}} \\
&\times \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}}
\end{aligned}
$$

La distribuzione congiunta condizionata delle tre variabili aleatorie a livello di *hashtag*, $(z^H_{udl}, h_{udl}, y^H_{udl})$, è data da:

$$
\begin{aligned}
&p(z^H_{udl}, h_{udl}, y^H_{udl}|x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\psi}_{1:T}, \pi^H, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&p(z^H_{udl}|\boldsymbol{\theta}_{ud}) p(h_{udl}, y^H_{udl}|z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \pi^H, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&p(z^H_{udl}|\boldsymbol{\theta}_{ud}) p(y^H_{udl}|\pi^H) p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^H_{udl}}} (\pi^H)^{y^H_{udl}} (1-\pi^H)^{1-y^H_{udl}} \\
&\times \prod_{h=1}^{H} \psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=1}} \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=0}} (\psi^{\mathcal{B}}_h)^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}}
\end{aligned}
$$

Assumendo che le parole in un documento non influenzino gli *hashtag* e viceversa, la distribuzione congiunta condizionata delle variabili a livello di parola e di *hashtag*, $(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}, \mathbf{z}^H_{ud}, \mathbf{h}_{ud}, \mathbf{y}^H_{ud})$, è data semplicemente dal

prodotto delle due distribuzioni precedenti:

$$\begin{aligned}
&p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}, \mathbf{z}^H_{ud}, \mathbf{h}_{ud}, \mathbf{y}^H_{ud} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&p(z^V_{udn}, w_{udn}, y^V_{udn} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(z^H_{udl}, h_{udl}, y^H_{udl} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\psi}_{1:T}, \pi^H, \boldsymbol{\psi}^{\mathcal{B}})
\end{aligned}$$

Data l'assunzione di scambiabilità delle parole e degli *hashtag* all'interno di un documento, la distribuzione congiunta condizionata delle variabili relative a un singolo documento è data da:

$$\begin{aligned}
&p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}, \mathbf{z}^H_{ud}, \mathbf{h}_{ud}, \mathbf{y}^H_{ud}, x_{ud}, z^*_{ud}, \boldsymbol{\lambda}_{ud}, \boldsymbol{\theta}_{ud} | \dots) \\
=&p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times p(\mathbf{z}^V_{ud}, \mathbf{w}_{ud}, \mathbf{y}^V_{ud}, \mathbf{z}^H_{ud}, \mathbf{h}_{ud}, \mathbf{y}^H_{ud} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times p(z^V_{udn}, w_{udn}, y^V_{udn} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(z^H_{udl}, h_{udl}, y^H_{udl} | x_{ud}, z^*_{ud}, \boldsymbol{\theta}_{ud}, \boldsymbol{\psi}_{1:T}, \pi^H, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) p(\boldsymbol{\lambda}_{ud}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times \prod_{l=1}^{L_{ud}} p(z^H_{udl}|\boldsymbol{\theta}_{ud}) p(y^H_{udl}|\pi^H) p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}})
\end{aligned}$$

dove $|\dots$ è un'abbreviazione di $|\boldsymbol{\theta}^*_u, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \pi^T_u, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}, \boldsymbol{\alpha}, \alpha_0$. Data l'assunzione di scambiabilità dei documenti scritti da uno stesso utente, la distribuzione congiunta condizionata dell'intera collezione è data da:

$$\begin{aligned}
&p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D} | \dots) \\
=&\prod_{u=1}^{U} \prod_{d=1}^{D_u} p(x_{ud}|\pi^T_u) p(z^*_{ud}|\boldsymbol{\theta}^*_u) \prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(y^V_{udn}|\pi^V) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times \prod_{l=1}^{L_{ud}} p(z^H_{udl}|\boldsymbol{\theta}_{ud}) p(y^H_{udl}|\pi^H) p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) \\
=&\prod_{u=1}^{U} \prod_{d=1}^{D_u} \left((\pi^T_u)^{x_{ud}} (1 - \pi^T_u)^{1 - x_{ud}} \right) \left(\prod_{t=1}^{T} (\theta^*_{u,t})^{\mathbb{1}_{z^*_{ud}=t}} \right)
\end{aligned}$$

$$
\begin{aligned}
&\times \left(\prod_{t=1}^{T} \delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}}\right)\left(\frac{\Gamma(\sum_{t=1}^{T}(\alpha_0+\lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t)}\prod_{t=1}^{T}\theta_{ud,t}^{\alpha_0+\lambda_{ud,t}\alpha_t-1}\right)\\
&\times \prod_{n=1}^{N_{ud}}\left(\prod_{t=1}^{T}\theta_{ud,t}^{\mathbb{1}_{z^V_{udn}}}(\pi^V)^{y^V_{udn}}(1-\pi^V)^{1-y^V_{udn}}\right.\\
&\times \left.\prod_{v=1}^{V}\phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}}\phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}}\left(\phi^{\mathcal{B}}_v\right)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}}\right)\\
&\times \prod_{l=1}^{L_{ud}}\left(\prod_{t=1}^{T}\theta_{ud,t}^{\mathbb{1}_{z^H_{udl}}}(\pi^H)^{y^H_{udl}}(1-\pi^H)^{1-y^H_{udl}}\right.\\
&\times \left.\prod_{h=1}^{H}\psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=1}}\psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=0}}\left(\psi^{\mathcal{B}}_h\right)^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}}\right)
\end{aligned}
$$

dove $|\ldots$ è un'abbreviazione di $|\boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}, \boldsymbol{\alpha}, \alpha_0$. Infine, siano

$$
\begin{aligned}
\mathbf{lat} &= (\mathbf{z}^V, \mathbf{y}^V, \mathbf{z}^H, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}),\\
\mathbf{par} &= (\boldsymbol{\theta}_{1:D}, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}),\\
\mathbf{oss} &= (\mathbf{w}, \mathbf{h}).
\end{aligned}
$$

La distribuzione congiunta delle variabili osservate, **oss**, e latenti, (**lat**, **par**), dati i parametri fissati è data da:

$$
\begin{aligned}
&p(\mathbf{lat}, \mathbf{oss}, \mathbf{par}|\boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})\\
=&p(\mathbf{lat}, \mathbf{oss}, \boldsymbol{\theta}_{1:D}|\boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}, \boldsymbol{\alpha}, \alpha_0)\\
&\times p(\boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\alpha}^*, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})\\
=&\prod_{u=1}^{U}\prod_{d=1}^{D_u} p(x_{ud}|\pi^T_u)p(z^*_{ud}|\boldsymbol{\theta}^*_u)\prod_{t=1}^{T}p(\lambda_{ud,t}|\delta)p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0)\\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud})p(y^V_{udn}|\pi^V)p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}})\\
&\times \prod_{l=1}^{L_{ud}} p(z^H_{udl}|\boldsymbol{\theta}_{ud})p(y^H_{udl}|\pi^H)p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}})\\
&\times p(\boldsymbol{\theta}^*_{1:U}|\boldsymbol{\alpha}^*)p(\boldsymbol{\phi}_{1:T}|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}_{1:T}|\boldsymbol{\beta}^H)p(\boldsymbol{\pi}^T_{1:U}|b^T_1, b^T_2)\\
&\times p(\delta|b^\delta_1, b^\delta_2)p(\pi^V|b^V_1, b^V_2)p(\pi^H|b^H_1, b^H_2)p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H)\\
=&p(\delta|b^\delta_1, b^\delta_2)p(\pi^V|b^V_1, b^V_2)p(\pi^H|b^H_1, b^H_2)p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H)
\end{aligned}
$$

$$
\begin{aligned}
&\times \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}_t|\boldsymbol{\beta}^H) \prod_{u=1}^{U} p(\boldsymbol{\theta}^*_u|\boldsymbol{\alpha}^*)p(\pi^T_u|b^T_1, b^T_2) \\
&\times \prod_{d=1}^{D_u} p(x_{ud}|\pi^T_u)p(z^*_{ud}|\boldsymbol{\theta}^*_u)p(\boldsymbol{\lambda}_{ud}|\delta)p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud})p(y^V_{udn}|\pi^V)p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times \prod_{l=1}^{L_{ud}} p(z^H_{udl}|\boldsymbol{\theta}_{ud})p(y^H_{udl}|\pi^H)p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) \\
= &\left(\frac{\Gamma(b^\delta_1 + b^\delta_2)}{\Gamma(b^\delta_1)\Gamma(b^\delta_2)}\delta^{b^\delta_1 - 1}(1-\delta)^{b^\delta_2 - 1}\right) \\
&\times \left(\frac{\Gamma(b^V_1 + b^V_2)}{\Gamma(b^V_1)\Gamma(b^V_2)}(\pi^V)^{b^V_1 - 1}(1-\pi^V)^{b^V_2 - 1}\right) \\
&\times \left(\frac{\Gamma(b^H_1 + b^H_2)}{\Gamma(b^H_1)\Gamma(b^H_2)}(\pi^H)^{b^H_1 - 1}(1-\pi^H)^{b^H_2 - 1}\right) \\
&\times \left(\frac{\Gamma(\sum_{v=1}^V \beta^V_v)}{\prod_{v=1}^V \Gamma(\beta^V_v)} \prod_{v=1}^{V} (\phi^{\mathcal{B}}_v)^{\beta^V_v - 1}\right) \left(\frac{\Gamma(\sum_{h=1}^H \beta^H_h)}{\prod_{h=1}^H \Gamma(\beta^H_h)} \prod_{h=1}^{H} (\psi^{\mathcal{B}}_h)^{\beta^H_h - 1}\right) \\
&\times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^V \boldsymbol{\beta}^V_v)}{\prod_{v=1}^V \Gamma(\beta^V_v)} \prod_{v=1}^{V} \phi_{t,v}^{\beta^V_v - 1}\right) \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{h=1}^H \boldsymbol{\beta}^H_h)}{\prod_{h=1}^H \Gamma(\beta^H_h)} \prod_{h=1}^{H} \psi_{t,h}^{\beta^H_h - 1}\right) \\
&\times \prod_{u=1}^{U} \left(\frac{\Gamma(\sum_{t=1}^T \alpha^*_t)}{\prod_{t=1}^T \Gamma(\alpha^*_t)} \prod_{t=1}^{T} (\theta^*_{u,t})^{\alpha^*_t - 1}\right) \\
&\times \left(\frac{\Gamma(b^T_1 + b^T_2)}{\Gamma(b^T_1)\Gamma(b^T_2)}(\pi^T_u)^{b^T_1 - 1}(1-\pi^T_u)^{b^T_2 - 1}\right) \\
&\times \prod_{d=1}^{D_u} \left((\pi^T_u)^{x_{ud}}(1-\pi^T_u)^{1-x_{ud}}\right) \left(\prod_{t=1}^{T} (\theta^*_{u,t})^{\mathbb{1}_{z^*_{ud}=t}}\right) \\
&\times \left(\prod_{t=1}^{T} \delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}}\right) \left(\frac{\Gamma(\sum_{t=1}^T (\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^T \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1}_{ud,t}\right) \\
&\times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta^{\mathbb{1}_{z^V_{udn}}}_{ud,t} (\pi^V)^{y^V_{udn}}(1-\pi^V)^{1-y^V_{udn}}\right. \\
&\times \left.\prod_{v=1}^{V} \phi^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}}_{z^V_{udn},v} \phi^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}}_{z^*_{ud},v} (\phi^{\mathcal{B}}_v)^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}}\right) \\
&\times \prod_{l=1}^{L_{ud}} \left(\prod_{t=1}^{T} \theta^{\mathbb{1}_{z^H_{udl}}}_{ud,t} (\pi^H)^{y^H_{udl}}(1-\pi^H)^{1-y^H_{udl}}\right.
\end{aligned}
$$

$$\times \prod_{h=1}^{H} \psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=1}} \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=0}} (\psi^{\mathcal{B}}_h)^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}} \Bigg)$$

In questo caso, si utilizzano le abbreviazioni **lat**, **oss** e **par** solo per rendere la notazione dell'ultima distribuzione più compatta: questa divisione in variabili latenti, osservate e parametri con una distribuzione a priori sarà utile nel Capitolo 4 per la costruzione del *Collapsed Gibbs Sampler*.

3.6 Relazione con Altri Modelli

Il modello proposto può essere visto come un'estensione della *Latent Dirichlet Allocation*, di *Twitter-LDA* e *Hashtag-LDA*: si mostra qui di seguito come è possibile passare dal modello proposto a uno dei suoi casi particolari fissando alcune variabili latenti. In particolare, si può ottenere la distribuzione congiunta delle variabili osservate e latenti del caso particolare –detto modello ristretto– a partire da quella del caso più generale –modello esteso– sfruttando il *teorema di Bayes*.

Siano A le variabili presenti solo nel modello ristretto e B quelle presenti nel modello esteso ma non in quello ristretto, allora la distribuzione del modello ristretto, p_{ristr}, può essere ottenuta come

$$p_{ristr} = \frac{p(A,B)}{p(B)}$$

Se si ha che tutte le distribuzioni delle variabili in B dipendono solo da altre variabili in B, allora vale

$$p_{ristr} = \frac{p(A,B)}{p(B)} = \frac{p(A|B)p(B)}{p(B)} = p(A|B)$$

Nel contesto dei *modelli generativi probabilistici* si è in questa situazione quando tutti i predecessori delle variabili aleatorie in B sono anch'essi in B; si consideri ad esempio il modello proposto in questa tesi: se si vuole assumere nota la variabile z^V_{udn}, allora è necessario assumere note anche $\boldsymbol{\theta}_{ud}$, $\boldsymbol{\lambda}_{ud}$ e δ poiché

- la distribuzione di z^V_{udn} dipende da $\boldsymbol{\theta}_{ud}$;
- la distribuzione di $\boldsymbol{\theta}_{ud}$ dipende da parametri fissati e $\boldsymbol{\lambda}_{ud}$;
- la distribuzione di $\boldsymbol{\lambda}_{ud}$ dipende da δ;

- la distribuzione di δ dipende solo da parametri fissati.

Infine, selezionando valori opportuni delle variabili in B, si ha che $p(A|B)$ coincide con la distribuzione congiunta del modello ristretto, p_{ristr}. In Tabella 3.2 si riportano le variabili latenti in B nei tre casi.

Di seguito si espone il procedimento completo per la *Latent Dirichlet Allocation*, mentre la derivazione di $p(A|B)$ viene omessa per gli altri due *topic model* poiché segue essenzialmente gli stessi passaggi.

topic model	$\mathbf{z}^V$	$\mathbf{y}^V$	$\mathbf{z}^H$	$\mathbf{y}^H$	$\mathbf{x}$	$\mathbf{z}^*$	$\boldsymbol{\lambda}$
LDA	×						
Twitter-LDA		×				×	
Hashtag-LDA				×		×	

Tabella 3.2: Lista delle variabili latenti presenti in *LDA*, *Twitter-LDA* e *Hashtag-LDA*.

3.6.1 Latent Dirichlet Allocation

La *Latent Dirichlet Allocation* è un caso particolare del modello proposto in cui

- gli *hashtag* non vengono considerati, ovvero $L = 0$;
- tutti i documenti trattano di più topic, ovvero x_{ud} $\forall ud$;
- tutti i documenti trattano di tutti i topic, ovvero $\lambda_{ud,y} = 1$ $\forall udt$;
- tutte le parole sono generate a partire da un topic, ovvero $y^V_{udn} = 1$ $\forall udn$.

In Figura 3.7 è rappresentato il *modello grafico probabilistico* della *Latent Dirichlet Allocation* come caso particolare del modello proposto.

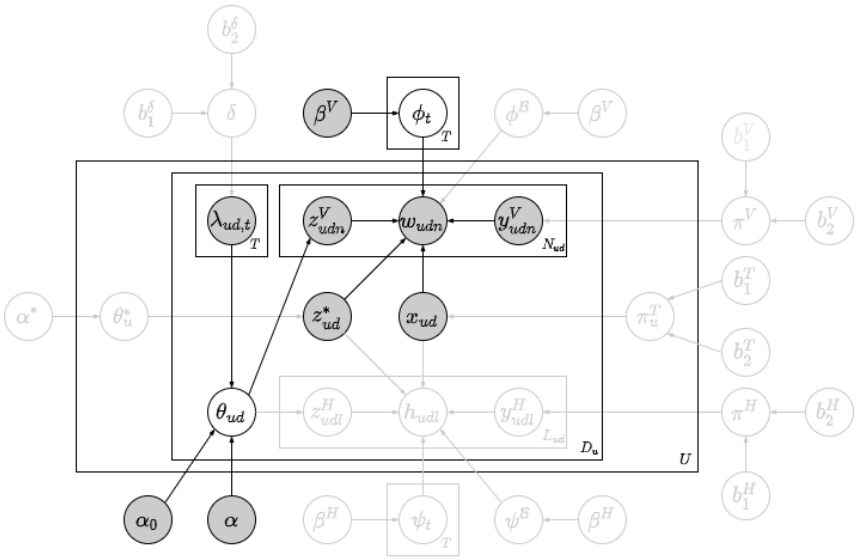


Figura 3.7: *Modello grafico probabilistico* della *LDA* come caso particolare del modello proposto.

Dal momento che non si considerano gli *hashtag*, tutte le variabili latenti legate ad essi sono rimosse dal modello; si ha quindi che

$$A = (\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T})$$
$$B = (\mathbf{y}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}})$$

Si vuole dimostrare che la distribuzione $p(A|B)$ coincide con la distribuzione congiunta della *Latent Dirichlet Allocation*

$$\begin{aligned}
p_{LDA} =& p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T} | \alpha, \beta^V) \\
=& \prod_{t=1}^{T} p(\boldsymbol{\phi}_t | \beta^V) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d | \alpha) \prod_{n=1}^{N_d} p(z^V_{dn} | \boldsymbol{\theta}_d) p(w_{dn} | z^V_{dn}, \boldsymbol{\phi}_{1:T}) \\
=& \left(\prod_{t=1}^{T} \frac{\Gamma(V\beta^V)}{\prod_{v=1}^{V} \Gamma(\beta^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta^V - 1} \right) \prod_{d=1}^{D} \left(\frac{\Gamma(T\alpha)}{\prod_{t=1}^{T} \Gamma(\alpha)} \prod_{t=1}^{T} \theta_{d,t}^{\alpha - 1} \right) \\
& \times \prod_{n=1}^{N_d} \left(\prod_{t=1}^{T} \theta_{d,t}^{\mathbb{1}_{z^V_{dn}=t}} \prod_{v=1}^{V} \phi_{z^V_{dn},v}^{\mathbb{1}_{w_{dn}=v}} \right)
\end{aligned}$$

La distribuzione di B può essere fattorizzata come segue

$$p(B) = p(\mathbf{y}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}})$$

$$
\begin{aligned}
&= p(\mathbf{y}^V|\pi^V)p(\pi^V) \times p(\mathbf{x}|\boldsymbol{\pi}^T_{1:U})p(\boldsymbol{\pi}^T_{1:U}) \\
&\quad \times p(\mathbf{z}^*|\boldsymbol{\theta}^*_{1:U})p(\boldsymbol{\theta}^*_{1:U}) \times p(\boldsymbol{\lambda}|\delta)p(\delta) \times p(\boldsymbol{\phi}^{\mathcal{B}})
\end{aligned}
$$

La distribuzione di (A, B) può essere fattorizzata come segue

$$
\begin{aligned}
p(A, B) =& p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T}, \mathbf{y}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^{\mathcal{B}}) \\
=& p(\mathbf{w}|\mathbf{y}^V, \mathbf{z}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \times p(\mathbf{y}^V|\pi^V)p(\pi^V) \\
&\times p(\mathbf{z}^V|\boldsymbol{\theta}_{1:D})p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\lambda}) \times p(\mathbf{x}|\boldsymbol{\pi}^T_{1:U})p(\boldsymbol{\pi}^T_{1:U}) \\
&\times p(\mathbf{z}^*|\boldsymbol{\theta}^*_{1:U})p(\boldsymbol{\theta}^*_{1:U}) \times p(\boldsymbol{\lambda}|\delta)p(\delta) \\
&\times p(\boldsymbol{\phi}_{1:T})p(\boldsymbol{\phi}^{\mathcal{B}})
\end{aligned}
$$

Riordinando i fattori, si può facilmente identificare $p(B)$:

$$
\begin{aligned}
=& p(\mathbf{w}|\mathbf{y}^V, \mathbf{z}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(\mathbf{z}^V|\boldsymbol{\theta}_{1:D})p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\lambda}) \times p(\boldsymbol{\phi}_{1:T}) \\
&\times p(\mathbf{y}^V|\pi^V)p(\pi^V) \times p(\mathbf{x}|\boldsymbol{\pi}^T_{1:U})p(\boldsymbol{\pi}^T_{1:U}) \\
&\times p(\mathbf{z}^*|\boldsymbol{\theta}^*_{1:U})p(\boldsymbol{\theta}^*_{1:U}) \times p(\boldsymbol{\lambda}|\delta)p(\delta) \times p(\boldsymbol{\phi}^{\mathcal{B}}) \\
=& p(\mathbf{w}|\mathbf{y}^V, \mathbf{z}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(\mathbf{z}^V|\boldsymbol{\theta}_{1:D})p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\lambda}) \times p(\boldsymbol{\phi}_{1:T}) \\
&\times p(B)
\end{aligned}
$$

Dividendo ambo i membri per $p(B)$[5], si ottiene

$$
\begin{aligned}
\frac{p(A, B)}{p(B)} =& p(\mathbf{w}|\mathbf{y}^V, \mathbf{z}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(\mathbf{z}^V|\boldsymbol{\theta}_{1:D})p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\lambda}) \times p(\boldsymbol{\phi}_{1:T})
\end{aligned}
$$

Quindi

$$
\begin{aligned}
p(A|B) =& \frac{p(A, B)}{p(B)} \\
=& p(\mathbf{w}|\mathbf{y}^V, \mathbf{z}^V, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&\times p(\mathbf{z}^V|\boldsymbol{\theta}_{1:D})p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\lambda}) \times p(\boldsymbol{\phi}_{1:T}) \\
=& \left(\prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \right)
\end{aligned}
$$

[5] Si dà per scontato che $p(B)$ sia maggiore di zero.

$$
\begin{aligned}
&\times \left(\prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) \right) \\
&\times \left(\prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \right) \\
=&\prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \\
&\times \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
=&\left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \left(\frac{\Gamma(\sum_{t=1}^{T} (\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1} \right) \\
&\times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{x_{ud}=1}} \right. \\
&\left. \times \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{x_{ud}=0}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1}} \right)
\end{aligned}
$$

Considerando $x_{ud} = 1$ per ogni documento ud, $\lambda_{ud,t} = 1$ per ogni topic t in ogni documento ud e $y^V_{udn} = 1$ per ogni parola udn, si ottiene

$$
\begin{aligned}
=&\left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \left(\frac{\Gamma(\sum_{t=1}^{T} (\alpha_0 + \alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_0 + \alpha_t - 1} \right) \\
&\times \prod_{n=1}^{N_{ud}} \left(\prod_{t=1}^{T} \theta_{ud,t}^{\mathbb{1}_{z^V_{udn}}} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}} \right)
\end{aligned}
$$

Infine, fissando $\alpha = \alpha_0 + \alpha_1 = \ldots = \alpha_0 + \alpha_T$ e $\beta^V = \beta_1^V = \ldots = \beta_V^V$, si può concludere che $p(A|B)$ coincide con p_{LDA}, quindi la *Latent Dirichlet Allocation* è un caso particolare del modello proposto.

3.6.2 Twitter-LDA

Twitter-LDA è un caso particolare del modello proposto in cui

- gli *hashtag* non vengono considerati, ovvero $L = 0$;
- tutti i documenti trattano di un unico topic, ovvero $x_{ud} = 0 \; \forall ud$.

In Figura 3.8 è rappresentato il *modello grafico probabilistico* di *Twitter-LDA* come caso particolare del modello proposto.

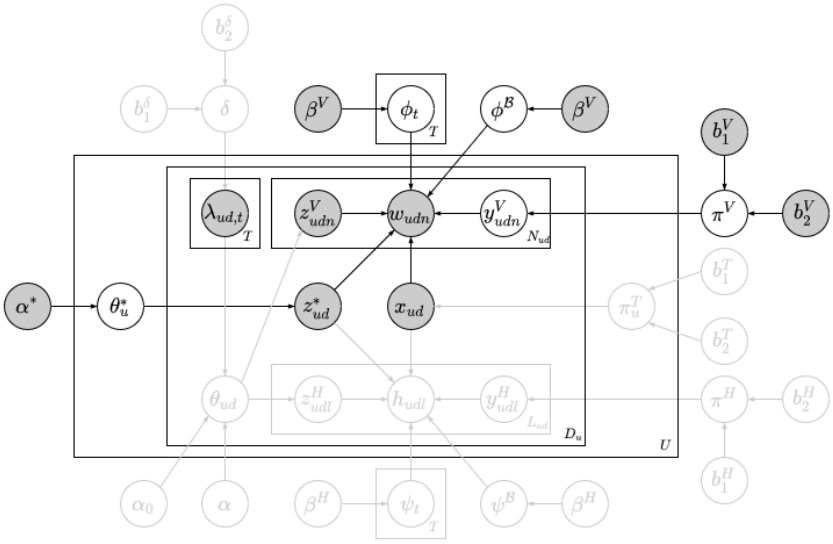


Figura 3.8: *Modello grafico probabilistico* di *Twitter-LDA* come caso particolare del modello proposto.

Dal momento che non si considerano gli *hashtag*, tutte le variabili latenti legate ad essi sono rimosse dal modello; si ha quindi che

$$A = (\mathbf{w}, \mathbf{y}^V, \mathbf{z}^*, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \pi^V, \boldsymbol{\phi}^{\mathcal{B}})$$
$$B = (\mathbf{z}^V, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\pi}^T_{1:U}, \delta)$$

da cui

$$\begin{aligned} p(A|B) = &\left(\frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)} (\pi^V)^{b_1^V - 1} (1 - \pi^V)^{b_2^V - 1} \right) \\ &\times \left(\frac{\Gamma(\sum_{v=1}^V \beta_v^V)}{\prod_{v=1}^V \Gamma(\beta_v^V)} \prod_{v=1}^V (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} \right) \\ &\times \left(\prod_{t=1}^T \frac{\Gamma(\sum_{v=1}^V \beta_v^V)}{\prod_{v=1}^V \Gamma(\beta_v^V)} \prod_{v=1}^V \phi_{t,v}^{\beta_v^V - 1} \right) \end{aligned}$$

$$
\begin{aligned}
&\times \prod_{u=1}^{U} \left(\frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \right) \\
&\times \prod_{d=1}^{D_u} \left(\prod_{t=1}^{T} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^* = t}} \right) \\
&\times \prod_{n=1}^{N_{ud}} \Big((\pi^V)^{y_{udn}^V} (1 - \pi^V)^{1 - y_{udn}^V} \\
&\times \prod_{v=1}^{V} \phi_{z_{udn}^V, v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 1} \mathbb{1}_{x_{ud}=1}} \\
&\times \phi_{z_{ud}^*, v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 1} \mathbb{1}_{x_{ud}=0}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 0}} \Big)
\end{aligned}
$$

Considerando $x_{ud} = 0$ per ogni documento ud, si ottiene

$$
\begin{aligned}
&= \left(\frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)} (\pi^V)^{b_1^V - 1} (1 - \pi^V)^{b_2^V - 1} \right) \\
&\times \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} \right) \\
&\times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\times \prod_{u=1}^{U} \left(\frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \right) \\
&\times \prod_{d=1}^{D_u} \left(\prod_{t=1}^{T} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^* = t}} \right) \\
&\times \prod_{n=1}^{N_{ud}} \Big((\pi^V)^{y_{udn}^V} (1 - \pi^V)^{1 - y_{udn}^V} \\
&\times \prod_{v=1}^{V} \phi_{z_{udn}^V, v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 1}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 0}} \Big)
\end{aligned}
$$

Infine, fissando $\alpha^* = \alpha_1^* = \ldots = \alpha_T^*$, $\beta^V = \beta_1^V = \ldots = \beta_V^V$ e $b^V = b_1^V = b_2^V$, si ottiene la distribuzione congiunta delle variabili osservate e latenti di *Twitter-LDA*. Si noti che questa distribuzione non è riportata in W. X. Zhao et al., 2011.

3.6.3 Hashtag-LDA

Hashtag-LDA è un caso particolare del modello proposto in cui

- tutti i documenti trattano di un unico topic, ovvero $x_{ud} = 0$ per ogni documento ud;
- tutte le parole sono generate a partire da un topic, ovvero $y^V_{udn} = 1$ per ogni parola udn.

In Figura 3.9 è rappresentato il *modello grafico probabilistico* di *Hashtag-LDA* come caso particolare del modello proposto.

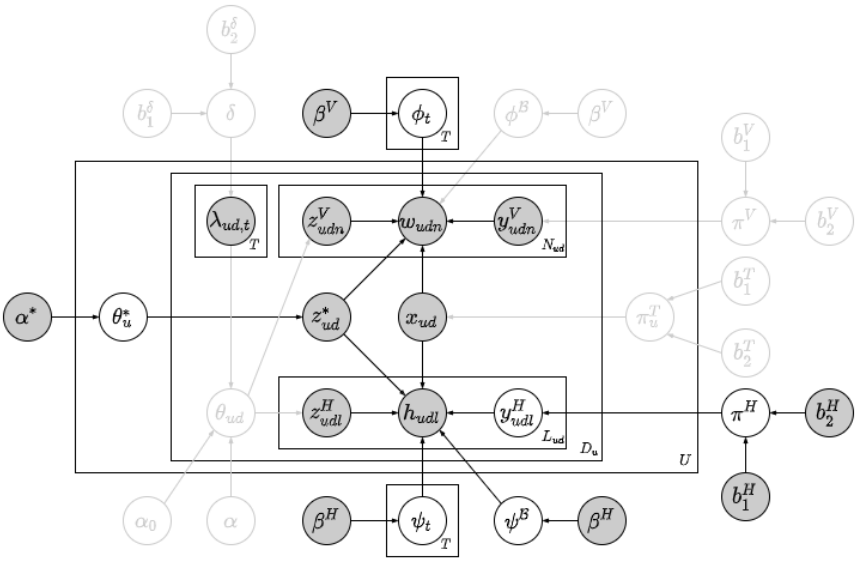


Figura 3.9: *Modello grafico probabilistico* di *Hashtag-LDA* come caso particolare del modello proposto.

In quest'ultimo caso si considerano sia le parole sia gli *hashtag*; si ha quindi che

$$A = (\mathbf{w}, \mathbf{h}, \mathbf{y}^H, \mathbf{z}^*, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \pi^H, \boldsymbol{\psi}^\mathcal{B})$$
$$B = (\mathbf{z}^V, \mathbf{y}^V, \mathbf{z}^H, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \boldsymbol{\phi}^\mathcal{B})$$

da cui

$$p(A|B) = \left(\frac{\Gamma(b_1^H + b_2^H)}{\Gamma(b_1^H)\Gamma(b_2^H)} (\pi^H)^{b_1^H - 1} (1 - \pi^H)^{b_2^H - 1} \right)$$

$$
\begin{aligned}
&\times \left(\frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} \right) \\
&\times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{H} \psi_{t,h}^{\beta_h^H - 1} \right) \\
&\times \prod_{u=1}^{U} \left(\frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \right) \\
&\times \prod_{d=1}^{D_u} \left(\prod_{t=1}^{T} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^* = t}} \right) \\
&\times \prod_{n=1}^{N_{ud}} \left(\prod_{v=1}^{V} \phi_{z_{udn}^V, v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 1} \mathbb{1}_{x_{ud}=1}} \right. \\
&\left. \times \phi_{z_{ud}^*, v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 1} \mathbb{1}_{x_{ud}=0}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V = 0}} \right) \\
&\times \prod_{l=1}^{L_{ud}} \Big((\pi^H)^{y_{udl}^H} (1 - \pi^H)^{1 - y_{udl}^H} \\
&\times \prod_{h=1}^{H} \psi_{z_{udl}^H, h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H = 1} \mathbb{1}_{x_{ud}=1}} \\
&\left. \times \psi_{z_{ud}^*, h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H = 1} \mathbb{1}_{x_{ud}=0}} (\psi_h^{\mathcal{B}})^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H = 0}} \right)
\end{aligned}
$$

Considerando $x_{ud} = 0$ per ogni documento ud e $y_{udn}^V = 1$ per ogni parola udn, si ottiene

$$
\begin{aligned}
&= \left(\frac{\Gamma(b_1^H + b_2^H)}{\Gamma(b_1^H)\Gamma(b_2^H)} (\pi^H)^{b_1^H - 1} (1 - \pi^H)^{b_2^H - 1} \right) \\
&\times \left(\frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} \right) \\
&\times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \\
&\times \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{H} \psi_{t,h}^{\beta_h^H - 1} \right) \\
&\times \prod_{u=1}^{U} \left(\frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \right)
\end{aligned}
$$

$$
\begin{aligned}
&\times \prod_{d=1}^{D_u} \left(\prod_{t=1}^{T} (\theta^*_{u,t})^{\mathbb{1}_{z^*_{ud}=t}} \right) \\
&\times \prod_{n=1}^{N_{ud}} \left(\prod_{v=1}^{V} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}} \right) \\
&\times \prod_{l=1}^{L_{ud}} \Big((\pi^H)^{y^H_{udl}} (1-\pi^H)^{1-y^H_{udl}} \\
&\times \prod_{h=1}^{H} \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}} (\psi^{\mathcal{B}}_h)^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}} \Big)
\end{aligned}
$$

Infine, fissando $\alpha^* = \alpha^*_1 = \ldots = \alpha^*_T$, $\beta^V = \beta^V_1 = \ldots = \beta^V_V$, $\beta^H = \beta^H_1 = \ldots = \beta^H_V$ e $b^H = b^H_1 = b^H_2$, si ottiene la distribuzione congiunta delle variabili osservate e latenti di *Hashtag-LDA*. Si noti che anche in questo caso la distribuzione congiunta delle variabili osservate e latenti non è riportata nell'articolo.

In realtà, F. Zhao et al., 2016 utilizzano due parametri diversi per le distribuzioni a priori di $\boldsymbol{\psi}_{1:T}$ e $\boldsymbol{\psi}^{\mathcal{B}}$, tuttavia nelle applicazioni pratiche questi due parametri –β^H e $\beta^{\mathcal{B}}$– assumono lo stesso valore: per semplicità, nel modello proposto si è considerato un unico parametro, emulando quanto proposto in *Twitter-LDA*.

Capitolo 4

Inferenza tramite Collapsed Gibbs Sampler

Il problema inferenziale che deve essere risolto nei *topic model* è il calcolo della distribuzione congiunta a posteriori delle variabili latenti **lat** date le variabili osservate **oss**; essa può essere ottenuta a partire dalla distribuzione congiunta delle due quantità definite sfruttando il *teorema di Bayes*:

$$p(\mathbf{lat}|\mathbf{oss}) = \frac{p(\mathbf{lat}, \mathbf{oss})}{p(\mathbf{oss})} = \frac{p(\mathbf{lat}, \mathbf{oss})}{\int_{S_{\mathbf{lat}}} p(\mathbf{lat}, \mathbf{oss}) d\mathbf{lat}}$$

dove $S_{\mathbf{lat}}$ è il supporto della distribuzione congiunta delle variabili latenti; per semplicità, i parametri fissati delle distribuzioni a priori sono stati omessi dalla formula. Analogamente a quanto osservato per la *Latent Dirichlet Allocation*, non è possibile effettuare inferenza esatta sulla struttura latente del modello a causa del denominatore. Tra i i diversi algoritmi applicabili, tra cui *Variational Inference*, *Expectation-Maximization algorithm*, *Gibbs Sampling* e *Collapsed Gibbs Sampling*, si è optato per l'ultimo poiché è lo stesso approccio adottato da W. X. Zhao et al., 2011 e F. Zhao et al., 2016 per effettuare l'inferenza a posteriori approssimata rispettivamente di *Twitter-LDA* e *Hashtag-LDA*. Inoltre, un *Collapsed Gibbs Sampler* è relativamente semplice da costruire rispetto ad altri algoritmi più complessi, come ad esempio *Variational Inference*.

Nella sezione 4.1 si introducono i metodi *Markov Chain Monte Carlo* e si espone la teoria alla base degli algoritmi *Gibbs Sampler* e *Collapsed Gibbs Sampler*; nella sezione 4.2 si espone il procedimento per costruire il *Collapsed*

Gibbs Sampler utilizzato per effettuare l'inferenza a posteriori approssimata del *topic model* proposto in questa tesi; infine, nella sezione 4.3 si propone un *Blocked Collapsed Gibbs Sampler* come algoritmo alternativo.

4.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (*MCMC*) è un metodo basato sul campionamento ripetuto che permette di costruire una catena di Markov ergodica con distribuzione stazionaria uguale a una distribuzione a posteriori d'interesse, di cui non necessariamente si conosce la costante di normalizzazione. La catena di Markov simulata può essere considerata come un campione generato dalla sua distribuzione stazionaria, di conseguenza le stime delle variabili latenti su cui la distribuzione a posteriori è definita possono essere ottenute applicando alla catena le tecniche *Monte Carlo*.

L'idea alla base del *MCMC* è molto semplice e intuitiva: si definisce uno spazio degli stati per le variabili aleatorie di cui si vuole ottenere una stima[1] e l'algoritmo fornisce una strategia per esplorare questo spazio degli stati. Per *stato* si intende un'assegnazione di valori a tutte le variabili aleatorie e a ogni iterazione dell'algoritmo si passa da uno stato dello spazio degli stati a un altro.

Formalmente, siano $\mathbf{X}$ le variabili osservate, $\mathbf{Z}$ la struttura latente e $\boldsymbol{\theta}$ i parametri del modello[2], allora un metodo *MCMC* permette di generare una catena di Markov con la seguente distribuzione stazionaria

$$p(\mathbf{U}|\mathbf{X}) = p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X})$$

dove $\mathbf{U} = (\mathbf{Z}, \boldsymbol{\theta})$ è una variabile K-dimensionale composta dalle variabili aleatorie su cui si vuole fare inferenza.

Prima di applicare le tecniche *Monte Carlo* è necessario considerare i seguenti aspetti:

- l'esecuzione di un metodo *MCMC* passa attraverso un periodo di *burn-in* in cui la catena di Markov non ha ancora raggiunto la stabilità e gli

[1] In un contesto bayesiano, come quello dei *topic model*, sono le variabili aleatorie su cui la distribuzione a posteriori è definita.

[2] $\mathbf{X}$ corrisponde a **oss**, $\mathbf{Z}$ corrisponde a **lat** e $\boldsymbol{\theta}$ corrisponde a **par**; la nuova notazione è tratta da Cohen, 2019.

stati appartenenti a questa fase non sono generati dalla distribuzione a posteriori d'interesse: considerare questi stati potrebbe portare a stime *Monte Carlo* errate, di conseguenza un approccio molto comune è scartarli;

- le stime *Monte Carlo* richiedono un campione indipendente ed identicamente distribuito, tuttavia gli stati di una catena di Markov sono per costruzione dipendenti: sfruttando il fatto che stati lontani tendono ad essere meno correlati tra loro, un approccio molto semplice per mitigare il problema, detto *thinning*, è considerare uno stato ogni $m \in \mathbb{N}$.

Formalmente, combinando i due approcci, l'insieme degli indici degli stati che vengono effettivamente utilizzati per calcolare le stime *Monte Carlo* è dato da

$$\mathcal{M} = \{i \in \{1, \dots, I\} : i > B, i \bmod m = 0\}$$

dove I è il numero di iterazioni, ed il numero delle iterazioni di *burn-in* B e il *lag* m variano in base alla complessità del problema inferenziale.

4.1.1 Gibbs Sampler

Il *Gibbs Sampler* (Geman & Geman, 1984) è uno dei più comuni metodi *MCMC* utilizzati nell'ambito del *bayesian natural language processing* (Cohen, 2019); esso esplora lo spazio degli stati aggiornando un elemento U_k di $\mathbf{U}$ alla volta estraendo un valore u_k dalla *full conditional probability* di u_k, ovvero dalla distribuzione condizionata di u_k date tutte le altre variabili del modello:

$$p(U_k|\mathbf{U}_{-k}, \mathbf{X}).$$

dove a ogni U_k corrisponde una singola variabile aleatoria su cui si vuole fare inferenza. Per poter passare dallo stato $\mathbf{U}^{(i)}$ allo stato $\mathbf{U}^{(i+1)}$ della catena è necessario fissare un ordine di aggiornamento che influenza come sono definite le *full conditional probabilities*; infatti, i valori aggiornati sono subito utilizzati avendo che una *full conditional probability* dipende dal valore più recente disponibile per ogni variabile condizionante. Un esempio molto semplice, tratto da Bishop, 2006, che permette di cogliere facilmente il meccanismo del *Gibbs Sampler*, considera una distribuzione su tre variabi-

li, $p(u_1, u_2, u_3)$; sia $u_k^{(i)}$ il valore assunto da u_k nell'i-mo stato della catena, allora

1. $u_1^{(i+1)}$ è estratto dalla distribuzione condizionata $p(u_1|u_2^{(i)}, u_3^{(i)})$;
2. $u_2^{(i+1)}$ è estratto dalla distribuzione condizionata $p(u_2|u_1^{(i+1)}, u_3^{(i)})$;
3. $u_3^{(i+1)}$ è estratto dalla distribuzione condizionata $p(u_3|u_2^{(i+1)}, u_3^{(i+1)})$.

Si riporta di seguito l'algoritmo: si noti che è necessario generare uno stato iniziale $\mathbf{U}^{(0)}$ poiché il *Gibbs Sampler* si limita ad aggiornare iterativamente gli stati della catena di Markov.

Algoritmo 1 *Gibbs Sampler*

Input: Variabili osservate $\mathbf{X}$
Output: Catena di Markov $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(I)}$

```
Inizializzazione dello stato iniziale U^(0)
for i = 1, ..., I do
    for k = 1, ..., K do
        Estrazione di u_k^(i) da p(u_k | u_1^(i), ..., u_{k-1}^(i), u_{k+1}^(i-1), ..., u_K^(i-1), X)
    end for
end for
return U^(1), ..., U^(I)
```

Nei *modelli generativi probabilistici* l'ordine di aggiornamento solitamente coincide con l'ordine di comparsa delle variabili nel *processo generativo*; analogamente, si sfruttano le dipendenze tra variabili definite nel *processo generativo* anche per inizializzare il primo stato della catena.

Infine, è importante notare che la derivazione di una *full conditional probability*, $p(U_k|\mathbf{U}_{-k}, \mathbf{X})$, è relativamente semplice dal momento che è proporzionale alla distribuzione congiunta di $\mathbf{U}$ e $\mathbf{X}$:

$$\begin{aligned} p(U_k|\mathbf{U}_{-k}, \mathbf{X}) &= \frac{p(U_k, \mathbf{U}_{-k}, \mathbf{X})}{p(\mathbf{U}_{-k}, \mathbf{X})} \\ &= \frac{p(\mathbf{U}, \mathbf{X})}{p(\mathbf{U}_{-k}, \mathbf{X})} \\ &\propto p(\mathbf{U}, \mathbf{X}) \end{aligned}$$

dove $\mathbf{U} = (U_k, \mathbf{U}_{-k})$ e $p(\mathbf{U}_{-k}, \mathbf{X})$ è trascurabile poiché non dipende da U_k.

4.1.2 Collapsed Gibbs Sampler

Spesso l'obiettivo primario della procedura di inferenza non è la stima dei parametri $\boldsymbol{\theta}$, ma la struttura latente $\mathbf{Z}$; in questo caso è possibile velocizzare l'algoritmo ed accorciare il periodo di *burn-in* della catena considerando come distribuzione stazionaria una distribuzione a posteriori in cui i parametri sono stati marginalizzati:

$$p(\mathbf{U}|\mathbf{X}) = p(\mathbf{Z}|\mathbf{X}) = \int_{S_{\boldsymbol{\theta}}} p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta}.$$

Un *Gibbs Sampling* in cui si utilizza una distribuzione a posteriori marginalizzata è detto *Collapsed Gibbs Sampler*. Si noti che, per utilizzare questa versione dell'algoritmo, $p(\mathbf{U}|\mathbf{X})$ deve avere una forma analitica chiusa: in ambito bayesano, si è in questa situazione quando i parametri marginalizzati provengono da distribuzione a priori coniugate. Sempre sfruttando le distribuzioni a priori coniugate, è possibile ottenere stime dei parametri $\boldsymbol{\theta}$ a partire dalla struttura latente $\mathbf{Z}$ stimata col *Collapsed Gibbs Sampler*: questo aspetto verrà trattato più approfonditamente nella sottosezione 4.2.5.

In generale, il *Collapsed Gibbs Sampler* risulta preferibile rispetto al *Gibbs Sampler* in termini di efficienza poiché, dovendo effettuare meno aggiornamenti di variabili, ogni iterazione dell'algoritmo necessita di meno tempo per essere completata. Questa caratteristica è molto attrattiva nell'ambito del *bayesian natural language processing* dal momento che i modelli sono solitamente molto complessi e il principale problema della procedura di inferenza è il tempo necessario per ottenere delle stime attendibili (Cohen, 2019).

4.2 Costruzione del Collapsed Gibbs Sampler

Il procedimento seguito per costruire il *Collapsed Gibbs Sampler* è tratto da Resnik e Hardisty, 2010 in cui è spiegato passo passo come fare inferenza su problemi legati all'analisi dei testi attraverso modelli bayesiani. I passi, esposti dettagliatamente nelle sezioni successive, sono i seguenti:

1. definizione del problema inferenziale;
2. derivazione della distribuzione congiunta di variabili latenti, osservate e parametri con una distribuzione a priori associata;

3. semplificazione della distribuzione congiunta tramite marginalizzazione dei parametri;
4. derivazione delle *full conditional probabilities*;
5. derivazione delle stime dei parametri marginalizzati.

Si noti che i punti 3 e 5 sono necessari solo per la costruzione di un *Collapsed Gibbs Sampler*: sono omessi nel caso di un *Gibbs Sampler*.

4.2.1 Problema Inferenziale

Il problema inferenziale che deve essere risolto nel modello proposto in questa tesi è il calcolo della distribuzione congiunta delle variabili latenti **lat** e dei parametri con una distribuzione a priori associata **par** date le variabili osservate **oss**[3]:

$$\begin{aligned} & p(\mathbf{lat}, \mathbf{par} | \mathbf{oss}, \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\ =& \frac{p(\mathbf{lat}, \mathbf{par}, \mathbf{oss} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})}{p(\mathbf{oss} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})} \\ =& \frac{p(\mathbf{lat}, \mathbf{par}, \mathbf{oss} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})}{\int_{S_{\mathbf{lat}}} \int_{S_{\mathbf{par}}} p(\mathbf{lat}, \mathbf{par}, \mathbf{oss} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) d\mathbf{lat} d\mathbf{par}} \end{aligned}$$

dove S_X indica il supporto di una variabile aleatoria X e

$$\begin{aligned} \mathbf{lat} &= (\mathbf{z}^V, \mathbf{y}^V, \mathbf{z}^H, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}) \\ \mathbf{par} &= (\boldsymbol{\theta}_{1:D}, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}^T_{1:U}, \delta, \pi^V, \pi^H, \boldsymbol{\phi}^{\mathcal{B}}, \boldsymbol{\psi}^{\mathcal{B}}) \\ \mathbf{oss} &= (\mathbf{w}, \mathbf{h}) \end{aligned}$$

4.2.2 Distribuzione Congiunta del Modello

La distribuzione congiunta delle variabili latenti, **lat**, dei parametri con una distribuzione a priori associata, **par**, e delle variabili osservate, **oss**, dati i parametri fissati $\boldsymbol{\alpha}^*$, $\boldsymbol{\alpha}$, α_0, $\boldsymbol{\beta}^V$, $\boldsymbol{\beta}^H$, $\mathbf{b} = (b_1^T, b_2^T, b_1^\delta, b_2^\delta, b_1^V, b_2^V, b_1^H, b_2^H)$, detta p_{mod}, è data da:

$$\begin{aligned} p_{mod} =& p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \mathbf{par} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\ =& p(\delta | b_1^\delta, b_2^\delta) \times p(\pi^V | b_1^V, b_2^V) \times p(\pi^H | b_1^H, b_2^H) \times p(\boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\beta}^V) \times p(\boldsymbol{\psi}^{\mathcal{B}} | \boldsymbol{\beta}^H) \end{aligned}$$

[3]Non si utilizza la notazione tratta da Bishop, 2006 per non creare confusione con le variabili presenti nel modello.

$$\times \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V)p(\boldsymbol{\psi}_t|\boldsymbol{\beta}^H) \times \prod_{u=1}^{U} p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*)p(\pi_u^T|b_1^T, b_2^T)$$
$$\times \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T)p(z_{ud}^*|\boldsymbol{\theta}_u^*)p(\boldsymbol{\lambda}_{ud}|\delta)p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0)$$
$$\times \prod_{n=1}^{N_{ud}} p(z_{udn}^V|\boldsymbol{\theta}_{ud})p(y_{udn}^V|\pi^V)p(w_{udn}|y_{udn}^V, z_{udn}^V, x_{ud}, z_{ud}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}})$$
$$\times \prod_{l=1}^{L_{ud}} p(z_{udl}^H|\boldsymbol{\theta}_{ud})p(y_{udl}^H|\pi^H)p(h_{udl}|y_{udl}^H, z_{udl}^H, x_{ud}, z_{ud}^*, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}})$$

4.2.3 Marginalizzazione dei Parametri

La distribuzione p_{mod} può essere riscritta come una produttoria di otto blocchi tali che ognuno dei parametri in **par** appare in un unico blocco:

$$\begin{aligned}
&p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \mathbf{par}|\boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\
&= \left(p(\delta|b_1^\delta, b_2^\delta) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) \right) \\
&\quad \times \left(p(\pi^V|b_1^V, b_2^V) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{n=1}^{N_{ud}} p(y_{udn}^V|\pi^V) \right) \\
&\quad \times \left(p(\pi^H|b_1^H, b_2^H) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{l=1}^{L_{ud}} p(y_{udl}^H|\pi^H) \right) \\
&\quad \times \left(\prod_{u=1}^{U} p(\pi_u^T|b_1^T, b_2^T) \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T) \right) \\
&\quad \times \left(\prod_{u=1}^{U} p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*) \prod_{d=1}^{D_u} p(z_{ud}^*|\boldsymbol{\theta}_u^*) \right) \\
&\quad \times \left(\prod_{u=1}^{U}\prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z_{udn}^V|\boldsymbol{\theta}_{ud}) \prod_{l=1}^{L_{ud}} p(z_{udl}^H|\boldsymbol{\theta}_{ud}) \right) \\
&\quad \times \left(p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V) \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{n=1}^{N_{ud}} p(w_{udn}|y_{udn}^V, z_{udn}^V, x_{ud}, z_{ud}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \right) \\
&\quad \times \left(p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H) \prod_{t=1}^{T} p(\boldsymbol{\psi}_t|\boldsymbol{\beta}^H) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{l=1}^{L_{ud}} p(h_{udl}|y_{udl}^H, z_{udl}^H, x_{ud}, z_{ud}^*, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) \right)
\end{aligned}$$

Ad esempio, si osserva che i parametri $\boldsymbol{\psi}^{\mathcal{B}}$ e $\boldsymbol{\psi}_{1:T}$ compaiono solo nell'ultima riga. A partire da quest'ultima formulazione i parametri in **par** possono

essere facilmente marginalizzati come segue:

$$
\begin{aligned}
&p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\
&= \int_{S_{\mathbf{par}}} p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \mathbf{par} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) d\mathbf{par} \\
&= \left(\int_{S_\delta} p(\delta | b_1^\delta, b_2^\delta) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{t=1}^{T} p(\lambda_{ud,t} | \delta) d\delta \right) \\
&\quad \times \left(\int_{S_{\pi^V}} p(\pi^V | b_1^V, b_2^V) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(y_{udn}^V | \pi^V) d\pi^V \right) \\
&\quad \times \left(\int_{S_{\pi^H}} p(\pi^H | b_1^H, b_2^H) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{l=1}^{L_{ud}} p(y_{udl}^H | \pi^H) d\pi^H \right) \\
&\quad \times \left(\int_{S_{\pi_{1:U}^T}} \prod_{u=1}^{U} p(\pi_u^T | b_1^T, b_2^T) \prod_{d=1}^{D_u} p(x_{ud} | \pi_u^T) d\boldsymbol{\pi}_{1:U}^T \right) \\
&\quad \times \left(\int_{S_{\boldsymbol{\theta}_{1:U}^*}} \prod_{u=1}^{U} p(\boldsymbol{\theta}_u^* | \boldsymbol{\alpha}^*) \prod_{d=1}^{D_u} p(z_{ud}^* | \boldsymbol{\theta}_u^*) d\boldsymbol{\theta}_{1:U}^* \right) \\
&\quad \times \left(\int_{S_{\boldsymbol{\theta}_{1:D}}} \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud} | \boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z_{udn}^V | \boldsymbol{\theta}_{ud}) \prod_{l=1}^{L_{ud}} p(z_{udl}^H | \boldsymbol{\theta}_{ud}) d\boldsymbol{\theta}_{1:D} \right) \\
&\quad \times \left(\int_{S_{\boldsymbol{\phi}^{\mathcal{B}}}} \int_{S_{\boldsymbol{\phi}_{1:T}}} p(\boldsymbol{\phi}^{\mathcal{B}} | \boldsymbol{\beta}^V) \prod_{t=1}^{T} p(\boldsymbol{\phi}_t | \boldsymbol{\beta}^V) \right. \\
&\quad \left. \times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(w_{udn} | y_{udn}^V, z_{udn}^V, x_{ud}, z_{ud}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) d\boldsymbol{\phi}^{\mathcal{B}} d\boldsymbol{\phi}_{1:T} \right) \\
&\quad \times \left(\int_{S_{\boldsymbol{\psi}^{\mathcal{B}}}} \int_{S_{\boldsymbol{\psi}_{1:T}}} p(\boldsymbol{\psi}^{\mathcal{B}} | \boldsymbol{\beta}^H) \prod_{t=1}^{T} p(\boldsymbol{\psi}_t | \boldsymbol{\beta}^H) \right. \\
&\quad \left. \times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{l=1}^{L_{ud}} p(h_{udl} | y_{udl}^H, z_{udl}^H, x_{ud}, z_{ud}^*, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) d\boldsymbol{\psi}^{\mathcal{B}} d\boldsymbol{\psi}_{1:T} \right)
\end{aligned}
$$

Per semplicità, di seguito verrà considerato un blocco alla volta; l'approccio utilizzato è lo stesso per tutti i blocchi e si può riassumere nel seguente procedimento:

1. si semplificano tutte le quantità che dipendono solo dai parametri fissati $\boldsymbol{\alpha}^*$, $\boldsymbol{\alpha}$, α_0, $\boldsymbol{\beta}^V$, $\boldsymbol{\beta}^H$, $\mathbf{b} = (b_1^T, b_2^T, b_1^\delta, b_2^\delta, b_1^V, b_2^V, b_1^H, b_2^H)$;
2. si sfrutta la scelta di utilizzare distribuzioni a priori coniugate per ottenere i nuclei delle distribuzioni a posteriori dei parametri;

3. si sostituisce l'integrale del nucleo di una distribuzione sul suo intero supporto con la sua costante di normalizzazione.

Per il primo blocco tutti i passaggi sono descritti anche a parole, mentre sono dati per scontati nei rimanenti sette.

Primo blocco

$$\int_{S_\delta} p(\delta|b_1^\delta, b_2^\delta) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) d\delta$$
$$= \int_{S_\delta} \frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)} \delta^{b_1^\delta - 1}(1-\delta)^{b_2^\delta - 1} \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{t=1}^{T} \delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}} d\delta$$

Si semplifica $\frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)}$ poiché dipende solo dai parametri fissati b_1^δ e b_2^δ.

$$\propto \int_{S_\delta} \delta^{b_1^\delta - 1}(1-\delta)^{b_2^\delta - 1} \delta^{\sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1}} (1-\delta)^{\sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0}} d\delta$$

Si riscrive la quantità precedente in modo da ottenere il nucleo di una distribuzione Beta di parametri $b_1^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1}$ e $b_2^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0}$.

$$= \int_{S_\delta} \delta^{b_1^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1} - 1}(1-\delta)^{b_2^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0} - 1} d\delta$$

L'integrale del nucleo di una distribuzione sul suo intero supporto è pari alla costante di normalizzazione della distribuzione.

$$= \frac{\Gamma(b_1^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1})\Gamma(b_2^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0})}{\Gamma(b_1^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1} + b_2^\delta + \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0})}$$

Si introduce una notazione compatta per i conteggi $\sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1}$ e $\sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0}$.

$$= \frac{\Gamma(b_1^\delta + n_\lambda^1)\Gamma(b_2^\delta + n_\lambda^0)}{\Gamma(b_1^\delta + b_2^\delta + DT)}$$

dove $n_\lambda^1 = \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1}$ è il numero totale di topic attivi, $n_\lambda^0 = \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0}$ è il numero totale di topic non attivi e $DT = n_\lambda^1 + n_\lambda^0$ è il numero totale di topic della collezione.

Secondo blocco

$$
\begin{aligned}
&\int_{S_{\pi^V}} p(\pi^V|b_1^V, b_2^V) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(y_{udn}^V|\pi^V) d\pi^V \\
=&\int_{S_{\pi^V}} \frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)} (\pi^V)^{b_1^V - 1}(1-\pi^V)^{b_2^V - 1} \prod_{udn} (\pi^V)^{y_{udn}^V}(1-\pi^V)^{1-y_{udn}^V} d\pi^V \\
\propto&\int_{S_{\pi^V}} (\pi^V)^{b_1^V - 1}(1-\pi^V)^{b_2^V - 1}(\pi^V)^{\sum_{udn} \mathbb{1}_{y_{udn}^V=1}}(1-\pi^V)^{\sum_{udn} \mathbb{1}_{y_{udn}^V=0}} d\pi^V \\
=&\int_{S_{\pi^V}} (\pi^V)^{b_1^V + \sum_{udn} \mathbb{1}_{y_{udn}^V=1} - 1}(1-\pi^V)^{b_2^V + \sum_{udn} \mathbb{1}_{y_{udn}^V=0} - 1} d\pi^V \\
=&\frac{\Gamma(b_1^V + \sum_{udn} \mathbb{1}_{y_{udn}^V=1})\Gamma(b_2^V + \sum_{udn} \mathbb{1}_{y_{udn}^V=0})}{\Gamma(b_1^V + \sum_{udn} \mathbb{1}_{y_{udn}^V=1} + b_2^V + \sum_{udn} \mathbb{1}_{y_{udn}^V=0})} \\
=&\frac{\Gamma(b_1^V + n_{y^V}^1)\Gamma(b_2^V + n_{y^V}^0)}{\Gamma(b_1^V + b_2^V + N)}
\end{aligned}
$$

dove $n_{y^V}^1 = \sum_{udn} \mathbb{1}_{y_{udn}^V=1}$ è il numero di parole generate a partire da un topic, $n_{y^V}^0 = \sum_{udn} \mathbb{1}_{y_{udn}^V=0}$ è il numero di parole di sottofondo e $N = n_{y^V}^1 + n_{y^V}^0$ è il numero totale di parole nella collezione.

Terzo blocco

$$
\begin{aligned}
&\int_{S_{\pi^H}} p(\pi^H|b_1^V, b_2^H) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{l=1}^{L_{ud}} p(y_{udl}^H|\pi^H) d\pi^H \\
=&\int_{S_{\pi^H}} \frac{\Gamma(b_1^H + b_2^H)}{\Gamma(b_1^H)\Gamma(b_2^H)} (\pi^H)^{b_1^H - 1}(1-\pi^H)^{b_2^H - 1} \prod_{udl} (\pi^H)^{y_{udl}^H}(1-\pi^H)^{1-y_{udl}^H} d\pi^H \\
\propto&\int_{S_\pi^H} (\pi^V)^{b_1^H - 1}(1-\pi^V)^{b_2^H - 1}(\pi^H)^{\sum_{udl} \mathbb{1}_{y_{udl}^H=1}}(1-\pi^H)^{\sum_{udl} \mathbb{1}_{y_{udl}^H=0}} d\pi^H \\
=&\int_{S_{\pi^H}} (\pi^H)^{b_1^H + \sum_{udl} \mathbb{1}_{y_{udl}^H=1} - 1}(1-\pi^H)^{b_2^H + \sum_{udl} \mathbb{1}_{y_{udl}^H=0} - 1} d\pi^H \\
=&\frac{\Gamma(b_1^H + \sum_{udl} \mathbb{1}_{y_{udl}^H=1})\Gamma(b_2^H + \sum_{udl} \mathbb{1}_{y_{udl}^H=0})}{\Gamma(b_1^H + \sum_{udl} \mathbb{1}_{y_{udl}^H=1} + b_2^H + \sum_{udl} \mathbb{1}_{y_{udl}^H=0})} \\
=&\frac{\Gamma(b_1^H + n_{y^H}^1)\Gamma(b_2^H + n_{y^H}^0)}{\Gamma(b_1^H + b_2^H + L)}
\end{aligned}
$$

dove $n_{y^H}^1 = \sum_{udl} \mathbb{1}_{y_{udl}^H=1}$ è il numero di *hashtag* generati a partire da un topic, $n_{y^H}^0 = \sum_{udl} \mathbb{1}_{y_{udl}^H=0}$ è il numero di *hashtag* globali e $L = n_{y^H}^1 + n_{y^H}^0$ è il numero totale di *hashtag* nella collezione.

Quarto blocco

$$
\begin{aligned}
&\int_{S_{\pi^T_{1:U}}} \prod_{u=1}^{U} p(\pi_u^T|b_1^T, b_2^T) \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T) d\boldsymbol{\pi}^T_{1:U} \\
&= \prod_{u=1}^{U} \int_{S_{\pi_u^T}} p(\pi_u^T|b_1^T, b_2^T) \prod_{d=1}^{D_u} p(x_{ud}|\pi_u^T) d\pi_u^T \\
&= \prod_{u=1}^{U} \int_{S_{\pi_u^T}} \frac{\Gamma(b_1^T + b_2^T)}{\Gamma(b_1^T)\Gamma(b_2^T)} (\pi_u^T)^{b_1^T - 1} (1 - \pi_u^T)^{b_2^T - 1} \prod_{d=1}^{D_u} (\pi_u^T)^{x_{ud}} (1 - \pi_u^T)^{1 - x_{ud}} d\pi_u^T \\
&\propto \prod_{u=1}^{U} \int_{S_{\pi_u^T}} (\pi_u^T)^{b_1^T - 1} (1 - \pi_u^T)^{b_2^T - 1} (\pi_u^T)^{\sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1}} (1 - \pi_u^T)^{\sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0}} d\pi_u^T \\
&= \prod_{u=1}^{U} \int_{S_{\pi_u^T}} (\pi_u^T)^{b_1^T + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1} - 1} (1 - \pi_u^T)^{b_2^T + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0} - 1} d\pi_u^T \\
&= \prod_{u=1}^{U} \frac{\Gamma(b_1^T + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1}) \Gamma(b_2^T + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0})}{\Gamma(b_1^V + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1} + b_2^V + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0})} \\
&= \prod_{u=1}^{U} \frac{\Gamma(b_1^T + n^1_{x_u}) \Gamma(b_2^T + n^0_{x_u})}{\Gamma(b_1^T + b_2^T + D_u)}
\end{aligned}
$$

dove $n^1_{x_u} = \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1}$ è il numero di documenti con più topic scritti dall'utente u, $n^0_{x_u} = \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0}$ è il numero di documenti con un unico topic scritti dall'utente u e $D_u = n^1_{x_u} + n^0_{x_u}$ è il numero totale di documenti scritti dall'utente u.

Quinto blocco

$$
\begin{aligned}
&\int_{S_{\boldsymbol{\theta}^*_{1:U}}} \prod_{u=1}^{U} p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*) \prod_{d=1}^{D_u} p(z_{ud}^*|\boldsymbol{\theta}_u^*) d\boldsymbol{\theta}^*_{1:U} \\
&= \prod_{u=1}^{U} \int_{S_{\boldsymbol{\theta}_u^*}} p(\boldsymbol{\theta}_u^*|\boldsymbol{\alpha}^*) \prod_{d=1}^{D_u} p(z_{ud}^*|\boldsymbol{\theta}_u^*) d\boldsymbol{\theta}_u^* \\
&= \prod_{u=1}^{U} \int_{S_{\boldsymbol{\theta}_u^*}} \frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \prod_{d=1}^{D_u} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^*=t}} d\boldsymbol{\theta}_u^* \\
&\propto \prod_{u=1}^{U} \int_{S_{\boldsymbol{\theta}_u^*}} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} (\theta_{u,t}^*)^{\sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^*=t}} d\boldsymbol{\theta}_u^* \\
&= \prod_{u=1}^{U} \int_{S_{\boldsymbol{\theta}_u^*}} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* + \sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^*=t} - 1} d\boldsymbol{\theta}_u^*
\end{aligned}
$$

$$= \prod_{u=1}^{U} \frac{\prod_{t=1}^{T} \Gamma(\alpha_t^* + \sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^*=t})}{\Gamma(\sum_{t=1}^{T}(\alpha_t^* + \sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^*=t}))}$$
$$= \prod_{u=1}^{U} \frac{\prod_{t=1}^{T} \Gamma(\alpha_t^* + n_{z_u^*}^t)}{\Gamma(\sum_{t=1}^{T} \alpha_t^* + D_u)}$$

dove $n_{z_u^*}^t = \sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^*=t}$ è il numero di documenti scritti dall'utente u il cui topic principale è t e $D_u = \sum_{t=1}^{T} n_{z_u^*}^t$ è il numero totale di documenti scritti dall'utente u.

Sesto blocco

$$\int_{S_{\boldsymbol{\theta}_{1:D}}} \prod_{u=1}^{U} \prod_{d=1}^{D_u} p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z_{udn}^V|\boldsymbol{\theta}_{ud}) \prod_{l=1}^{L_{ud}} p(z_{udl}^H|\boldsymbol{\theta}_{ud}) d\boldsymbol{\theta}_{1:D}$$
$$= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \int_{S_{\boldsymbol{\theta}_{ud}}} p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z_{udn}^V|\boldsymbol{\theta}_{ud}) \prod_{l=1}^{L_{ud}} p(z_{udl}^H|\boldsymbol{\theta}_{ud}) d\boldsymbol{\theta}_{ud}$$
$$= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \int_{S_{\boldsymbol{\theta}_{ud}}} \frac{\Gamma(\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1} \prod_{n=1}^{N_{ud}} \theta_{ud,t}^{\mathbb{1}_{z_{udn}^V=t}} \prod_{l=1}^{L_{ud}} \theta_{ud,t}^{\mathbb{1}_{z_{udl}^H=t}} d\boldsymbol{\theta}_{ud}$$
$$= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \frac{\Gamma(\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \int_{S_{\boldsymbol{\theta}_{ud}}} \prod_{t=1}^{T} \theta_{ud,t}^{\alpha_0 + \lambda_{ud,t}\alpha_t + \sum_{n=1}^{N_{ud}} \mathbb{1}_{z_{udn}^V=t} + \sum_{l=1}^{L_{ud}} \mathbb{1}_{z_{udl}^H=t} - 1} d\boldsymbol{\theta}_{ud}$$
$$= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \frac{\Gamma(\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \frac{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t + \sum_{n=1}^{N_{ud}} \mathbb{1}_{z_{udn}^V=t} + \sum_{l=1}^{L_{ud}} \mathbb{1}_{z_{udl}^H=t})}{\Gamma(\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t + \sum_{n=1}^{N_{ud}} \mathbb{1}_{z_{udn}^V=t} + \sum_{l=1}^{L_{ud}} \mathbb{1}_{z_{udl}^H=t}))}$$
$$= \prod_{u=1}^{U} \prod_{d=1}^{D_u} \frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T} \lambda_{ud,t}\alpha_t)}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \frac{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t + n_{z_{ud}^V, z_{ud}^H}^t)}{\Gamma(T\alpha_0 + \sum_{t=1}^{T} \lambda_{ud,t}\alpha_t + N_{ud} + L_{ud})}$$

dove $n_{z_{ud}^V, z_{ud}^H}^t = \sum_{n=1}^{N_{ud}} \mathbb{1}_{z_{udn}^V=t} + \sum_{l=1}^{L_{ud}} \mathbb{1}_{z_{udl}^H=t}$ è il numero di elementi testuali –parole, *hashtag* o entrambi– del d-mo documento dell'utente u a cui è associato il topic t senza considerare né il tipo di ogni documento né l'origine di ogni elemento e $N_{ud} + L_{ud} = \sum_{t=1}^{T} n_{z_{ud}^V, z_{ud}^H}^t$ è il numero totale elementi testuali del d-mo documento dell'utente u.

Settimo blocco

$$\int_{S_{\boldsymbol{\phi}^{\mathcal{B}}}} \int_{S_{\boldsymbol{\phi}_{1:T}}} p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V) \prod_{t=1}^{T} p(\boldsymbol{\phi}_t|\boldsymbol{\beta}^V)$$
$$\times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(w_{udn}|y_{udn}^V, z_{udn}^V, x_{ud}, z_{ud}^*, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) d\boldsymbol{\phi}^{\mathcal{B}} d\boldsymbol{\phi}_{1:T}$$
$$= \int_{S_{\boldsymbol{\phi}^{\mathcal{B}}}} \int_{S_{\boldsymbol{\phi}_{1:T}}} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} \right) \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right)$$

$$\times \left(\prod_{udn} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \right) d\boldsymbol{\phi}^{\mathcal{B}} d\boldsymbol{\phi}_{1:T}$$

$$\propto \int_{S_{\boldsymbol{\phi}^{\mathcal{B}}}} \prod_{v=1}^{V} (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} (\phi_v^{\mathcal{B}})^{\sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} d\boldsymbol{\phi}^{\mathcal{B}} \times \int_{S_{\boldsymbol{\Phi}_{1:T}}} \left(\prod_{t=1}^{T} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right)$$

$$\times \left(\prod_{t=1}^{T} \prod_{udn} \prod_{v=1}^{V} \phi_{t,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1}} \phi_{t,v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0}} \right) d\boldsymbol{\phi}_{1:T}$$

$$\propto \int_{S_{\boldsymbol{\phi}^{\mathcal{B}}}} \prod_{v=1}^{V} (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} (\phi_v^{\mathcal{B}})^{\sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} d\boldsymbol{\phi}^{\mathcal{B}}$$

$$\times \prod_{t=1}^{T} \int_{S_{\boldsymbol{\phi}_t}} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \phi_{t,v}^{\sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0})} d\boldsymbol{\phi}_t$$

$$\propto \int_{S_{\boldsymbol{\phi}^{\mathcal{B}}}} \prod_{v=1}^{V} (\phi_v^{\mathcal{B}})^{\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0} - 1} d\boldsymbol{\phi}^{\mathcal{B}}$$

$$\times \prod_{t=1}^{T} \int_{S_{\boldsymbol{\phi}_t}} \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0}) - 1} d\boldsymbol{\phi}_t$$

$$= \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0})}{\Gamma(\sum_{v=1}^{V}(\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}))}$$

$$\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0}))}{\Gamma(\sum_{v=1}^{V}(\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0})))}$$

$$= \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^v_{y^V=0})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot}_{y^V=0})} \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot,t}_{y^V=1})}$$

dove $n^v_{y^V=0} = \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}$ è il numero di volte che v appare come parola di sottofondo e $n^{\cdot}_{y^V=0} = \sum_{v=1}^{V} n^v_{y^V=0}$ è il numero totale di parole di sottofondo nella collezione, $n^{v,t}_{y^V=1} = \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0})$ è il numero di volte la parola v è associata al topic t considerando sia il tipo di ogni documento sia l'origine di ogni parola e $n^{\cdot,t}_{y^V=1} = \sum_{v=1}^{V} n^{v,t}_{y^V=1}$ è il numero totale di parole associate al topic t considerando sia il tipo di ogni documento sia l'origine di ogni parola.

Ottavo blocco

$$\int_{S_{\boldsymbol{\psi}^{\mathcal{B}}}} \int_{S_{\boldsymbol{\psi}_{1:T}}} p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H) \prod_{t=1}^{T} p(\boldsymbol{\psi}_t|\boldsymbol{\beta}^H)$$

$$\times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{l=1}^{L_{ud}} p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) d\boldsymbol{\psi}^{\mathcal{B}} d\boldsymbol{\psi}_{1:T}$$

$$= \int_{S_{\boldsymbol{\psi}^{\mathcal{B}}}} \int_{S_{\boldsymbol{\psi}_{1:T}}} \left(\frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{V} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} \right) \left(\prod_{t=1}^{T} \frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{H} \psi_{t,v}^{\beta_h^H - 1} \right)$$

$$
\begin{aligned}
&\times \left(\prod_{udl} \prod_{h=1}^{H} \psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=1}} \phi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{x_{ud}=0}} (\psi_h^{\mathcal{B}})^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}} \right) d\boldsymbol{\psi}^{\mathcal{B}} d\boldsymbol{\psi}_{1:T} \\
\propto &\int_{S_{\boldsymbol{\psi}^{\mathcal{B}}}} \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} (\psi_h^{\mathcal{B}})^{\sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}} d\boldsymbol{\psi}^{\mathcal{B}} \times \int_{S_{\boldsymbol{\psi}_{1:T}}} \left(\prod_{t=1}^{T} \prod_{h=1}^{H} \psi_{t,h}^{\beta_h^H - 1} \right) \\
&\times \left(\prod_{t=1}^{T} \prod_{udl} \prod_{h=1}^{H} \psi_{t,h}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{z^H_{udl}=t}\mathbb{1}_{x_{ud}=1}} \psi_{t,v}^{\mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}\mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0}} \right) d\boldsymbol{\psi}_{1:T} \\
\propto &\int_{S_{\boldsymbol{\psi}^{\mathcal{B}}}} \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} (\psi_h^{\mathcal{B}})^{\sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}} d\boldsymbol{\psi}^{\mathcal{B}} \\
&\times \prod_{t=1}^{T} \int_{S_{\boldsymbol{\psi}_t}} \prod_{h=1}^{H} \psi_{t,h}^{\beta_h^H - 1} \psi_{t,v}^{\sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}(\mathbb{1}_{z^H_{udl}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0})} d\boldsymbol{\psi}_t \\
\propto &\int_{S_{\boldsymbol{\psi}^{\mathcal{B}}}} \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0} - 1} d\boldsymbol{\psi}^{\mathcal{B}} \\
&\times \prod_{t=1}^{T} \int_{S_{\boldsymbol{\psi}_t}} \prod_{h=1}^{H} \psi_{t,h}^{\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}(\mathbb{1}_{z^H_{udl}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0}) - 1} d\boldsymbol{\psi}_t \\
= &\frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0})}{\Gamma(\sum_{h=1}^{H}(\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}))} \\
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}(\mathbb{1}_{z^H_{udl}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0}))}{\Gamma(\sum_{h=1}^{H}(\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}(\mathbb{1}_{z^H_{udl}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0})))} \\
= &\frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n^h_{y^H=0})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n^{\cdot}_{y^H=0})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n^{\cdot,t}_{y^H=1})}
\end{aligned}
$$

dove $n^h_{y^H=0} = \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=0}$ è il numero di volte che h appare come *hashtag* globale e $n^{\cdot}_{y^H=0} = \sum_{h=1}^{H} n^h_{y^H=0}$ è il numero totale di *hashtag* globali nella collezione, $n^{h,t}_{y^H=1} = \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y^H_{udl}=1}(\mathbb{1}_{z^H_{udl}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}}\mathbb{1}_{x_{ud}=0})$ è il numero di volte l' *hashtag* h è associato al topic t considerando sia il tipo di ogni documento sia l'origine di ogni *hashtag* e $n^{\cdot,t}_{y^H=1} = \sum_{h=1}^{H} n^{h,t}_{y^H=1}$ è il numero totale di *hashtag* associati al topic t considerando sia il tipo di ogni documento sia l'origine di ogni *hashtag*.

Distribuzione congiunta marginalizzata

La distribuzione congiunta di $(\mathbf{lat}, \mathbf{oss})$ dati i parametri fissati è quindi data dalla produttoria di dieci blocchi, ognuno dei quali corrisponde a uno dei parametri in **par**:

$$
\begin{aligned}
&p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\
= &\int_{S_{\mathbf{par}}} p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda}, \mathbf{par} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) d\mathbf{par}
\end{aligned}
$$

$$\propto \left(\frac{\Gamma(b_1^\delta + n_\lambda^1)\Gamma(b_2^\delta + n_\lambda^0)}{\Gamma(b_1^\delta + b_2^\delta + DT)}\right) \tag{1}$$

$$\times \left(\frac{\Gamma(b_1^V + n_{y^V}^1)\Gamma(b_2^V + n_{y^V}^0)}{\Gamma(b_1^V + b_2^V + L)}\right) \tag{2}$$

$$\times \left(\frac{\Gamma(b_1^H + n_{y^H}^1)\Gamma(b_2^H + n_{y^H}^0)}{\Gamma(b_1^H + b_2^H + L)}\right) \tag{3}$$

$$\times \left(\prod_{u=1}^{U} \frac{\Gamma(b_1^T + n_{x_u}^1)\Gamma(b_2^T + n_{x_u}^0)}{\Gamma(b_1^T + b_2^T + D_u)}\right) \tag{4}$$

$$\times \left(\prod_{u=1}^{U} \frac{\prod_{t=1}^{T}\Gamma(\alpha_t^* + n_{z_u^*}^t)}{\Gamma(\sum_{t=1}^{T}\alpha_t^* + D_u)}\right) \tag{5}$$

$$\times \left(\prod_{u=1}^{U}\prod_{d=1}^{D_u} \frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ud,t}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \frac{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t + n_{z_{ud}^V, z_{ud}^H}^t)}{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ud,t}\alpha_t + N_{ud} + L_{ud})}\right) \tag{6}$$

$$\times \left(\frac{\prod_{v=1}^{V}\Gamma(\beta_v^V + n_{y^V=0}^v)}{\Gamma(\sum_{v=1}^{V}\beta_v^V + n_{y^V=0}^{\cdot})}\right) \tag{7}$$

$$\times \left(\prod_{t=1}^{T} \frac{\prod_{v=1}^{V}\Gamma(\beta_v^V + n_{y^V=1}^{v,t})}{\Gamma(\sum_{v=1}^{V}\beta_v^V + n_{y^V=1}^{\cdot,t})}\right) \tag{8}$$

$$\times \left(\frac{\prod_{h=1}^{H}\Gamma(\beta_h^H + n_{y^H=0}^h)}{\Gamma(\sum_{h=1}^{H}\beta_h^H + n_{y^H=0}^{\cdot})}\right) \tag{9}$$

$$\times \left(\prod_{t=1}^{T} \frac{\prod_{h=1}^{H}\Gamma(\beta_h^H + n_{y^H=1}^{h,t})}{\Gamma(\sum_{h=1}^{H}\beta_h^H + n_{y^H=1}^{\cdot,t})}\right) \tag{10}$$

4.2.4 Full Conditional Probabilities

La *full conditional probability* di ogni variabile latente è proporzionale alla distribuzione congiunta ottenuta marginalizzando le variabili latenti in **par**:

$$p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})$$

Si consideri ad esempio la *full conditional probability* di x_{ij}: sfruttando il *teorema di Bayes*, essa può essere calcolata come

$$\begin{aligned} &p(x_{ij} | \mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}_{-ij}, \mathbf{z}^*, \lambda, \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\ =&\frac{p(x_{ij}, \mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}_{-ij}, \mathbf{z}^*, \lambda | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})}{p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}_{-ij}, \mathbf{z}^*, \lambda | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})} \end{aligned}$$

Il denominatore può essere trascurato poiché non dipende da x_{ij}.

$$\propto p(x_{ij}, \mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}_{-ij}, \mathbf{z}^*, \lambda | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})$$

Dato che $(x_{ij}, \mathbf{x}_{-ij}) = \mathbf{x}$, il numeratore coincide con la distribuzione congiunta di variabili latenti, parametri e variabili osservate.

$$= p(\mathbf{z}^V, \mathbf{w}, \mathbf{y}^V, \mathbf{z}^H, \mathbf{h}, \mathbf{y}^H, \mathbf{x}, \mathbf{z}^*, \boldsymbol{\lambda} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b})$$

Il procedimento per ottenere le distribuzioni condizionate è essenzialmente sempre lo stesso e consiste in:

1. considerare solo i blocchi –per comodità sono stati numerati da (1) a (10)– in cui compare la variabile di cui si vuole calcolare la distribuzione;
2. identificare i conteggi all'interno delle funzioni Gamma e determinare quando la parte del conteggio relativa alla variabile di cui si vuole calcolare la distribuzione è non nulla;
3. applicare la proprietà della funzione Gamma
$$\Gamma(x + k) = \prod_{q=0}^{k-1} (x + q)\Gamma(x)$$
per $x > 0$ e $k \in \mathbb{N}$, in modo tale che tutte le funzioni Gamma non dipendano dalla variabili di cui si vuole calcolare la distribuzione;
4. trascurare le funzioni Gamma e considerare solo le quantità ottenute applicando la proprietà della funzione Gamma.

Si noti che dopo aver concluso il terzo passaggio si ottiene la produttoria di funzioni Gamma di partenza, in cui però gli argomenti non contengono i conteggi relativi alla variabile di cui si vuole calcolare la distribuzione, moltiplicata per un secondo blocco –composto dalle quantità ottenute applicando la proprietà della funzione Gamma– che costituisce la distribuzione finale d'interesse.

Per capire l'idea di base, si consideri la seguente quantità:

$$\frac{\Gamma(\alpha_1 + 1)\Gamma(\alpha_2)\Gamma(\alpha_3 + 5)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + 6)}$$

Applicando la proprietà della funzione Gamma, si ottiene

$$\frac{\alpha_1\Gamma(\alpha_1)\Gamma(\alpha_2)\prod_{q=0}^{5-1}(\alpha_3 + q)\Gamma(\alpha_3)}{\prod_{q=0}^{6-1}(\alpha_1 + \alpha_2 + \alpha_3 + q)\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}$$

Svolgendo i conti, si ottengono le funzioni Gamma di partenza, con argomenti privi dei conteggi, moltiplicate per un secondo blocco:

$$\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1+\alpha_2+\alpha_3)} \times \frac{\alpha_1 \prod_{q=0}^{4}(\alpha_3+q)}{\prod_{q=0}^{5}(\alpha_1+\alpha_2+\alpha_3+q)}$$

Di seguito, per rendere il procedimento meno pesante, si omette il passaggio in cui si isola la produttoria di funzioni Gamma che non dipendono dalla variabili d'interesse e si trascura direttamente il tutto. Si noti che queste quantità sono trascurabili poiché, non dipendendo dal valore assunto dalla variabile d'interesse, sono uguali per tutte le possibili modalità della variabile d'interesse.

Si noti inoltre che formalmente non è sbagliato applicare la proprietà anche a funzioni Gamma in cui l'argomento non ha un conteggio poiché si sfrutta la definizione di *prodotto vuoto*:

$$\Gamma(\alpha+0) = \prod_{q=0}^{0-1}(\alpha+q)\Gamma(\alpha) = \prod_{q=0}^{-1}(\alpha+q)\Gamma(\alpha) = \Gamma(\alpha)$$

Si specifica ciò poiché la maggior parte delle produttorie ha l'indice inferiore maggiore dell'indice superiore dal momento che il conteggio corrispondente a quest'ultimo risulta essere nullo.

Infine, è interessante notare che non è necessario rimuovere le funzioni Gamma dalle distribuzioni condizionate, tuttavia è fortemente consigliato – di fatto è un passaggio standard– poiché sono funzioni relativamente lente da calcolare e la loro rimozione permette di implementare algoritmi più efficienti.

Notazione

Siano ij gli indici di una variabile associata a un documento e ijk quelli di una variabile associata a un elemento testuale –parola o *hashtag*– di cui si vuole calcolare la *full conditional probability*; si assume che

t^* sia il topic principale del documento ij;
w^* sia la parola osservata in posizione k nel documento ij;
h^* sia l'*hashtag* osservato in posizione k nel documento ij;
t^V sia il topic associato alla parola in posizione k nel documento ij;
t^H sia il topic associato all'*hashtag* in posizione k nel documento ij

Visto che le distribuzioni delle sezioni successive sono tutte *full conditional probability*, si omette la parte dopo la barra verticale | e si scrive |... per rendere la notazione più compatta. Ad esempio, si consideri la distribuzione di x_{ij} date tutte le altre variabili, allora vale

$$p(x_{ij}|\ldots) = p(x_{ij}|(\mathbf{lat}, \mathbf{oss}, \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \setminus x_{ij})$$

Se il valore di una variabile fissata viene specificato, ciò significa che la variabile compare nella definizione della *full conditional probability* oppure, nel caso di x_{ij}, y^V_{ijk} e y^H_{ijk}, che si hanno diverse distribuzioni in base al valore assunto dalla variabile dicotomica.

Distribuzione del tipo del documento

La distribuzione di x^*_{ij} date tutte le altre variabili è proporzionale al prodotto dei blocchi (4), (8) e (10):

$$\begin{aligned}
&p(x_{ij}|\ldots) \\
&\propto \prod_{u=1}^{U} \frac{\Gamma(b_1^T + n^1_{x_u})\Gamma(b_2^T + n^0_{x_u})}{\Gamma(b_1^T + b_2^T + D_u)} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot,t}_{y^V=1})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n^{\cdot,t}_{y^H=1})} \\
&\propto \frac{\Gamma(b_1^T + n^1_{x_i})\Gamma(b_2^T + n^0_{x_i})}{\Gamma(b_1^T + b_2^T + D_i)} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot,t}_{y^V=1})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n^{\cdot,t}_{y^H=1})}
\end{aligned}$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare x_{ij}:

$$\begin{aligned}
n^1_{x_i} &= \sum_{d=1}^{D_i} \mathbb{1}_{x_{id}=1} = \sum_{n \neq j} \mathbb{1}_{x_{id}=1} + \mathbb{1}_{x_{ij}=1} \\
&= n^1_{x_i,-j} + \mathbb{1}_{x_{ij}=1} = \begin{cases} n^1_{x_i,-j} + 1 & \text{se } x_{ij} = 1 \\ n^1_{x_i,-j} & \text{se } x_{ij} = 0 \end{cases} \\
n^0_{x_i} &= \sum_{d=1}^{D_i} \mathbb{1}_{x_{id}=0} = \sum_{n \neq j} \mathbb{1}_{x_{id}=0} + \mathbb{1}_{x_{ij}=0}
\end{aligned}$$

$$= n^{0}_{x_i,-j} + \mathbb{1}_{x_{ij}=0} = \begin{cases} n^{1}_{x_i,-j} & \text{se } x_{ij} = 1 \\ n^{1}_{x_i,-j} + 1 & \text{se } x_{ij} = 0 \end{cases}$$

$$n^{v,t}_{y^V=1} = \sum_{u=1}^{U}\sum_{d=1}^{D_u}\sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t}\mathbb{1}_{x_{ud}=0})$$

$$= n^{v,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}(\mathbb{1}_{z^V_{ijn}=t}\mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0})$$

$$= \begin{cases} n^{v,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}\mathbb{1}_{z^V_{ijn}=t} & \text{se } x_{ij} = 1 \\ n^{v,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} = t \\ n^{v,t}_{y^V=1,-ij} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} \neq t \end{cases}$$

$$n^{\cdot,t}_{y^V=1} = \sum_{v=1}^{V} n^{v,t}_{y^V=1} = n^{\cdot,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1}(\mathbb{1}_{z^V_{ijn}=t}\mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0})$$

$$= \begin{cases} n^{\cdot,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1}\mathbb{1}_{z^V_{ijn}=t} & \text{se } x_{ij} = 1 \\ n^{\cdot,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} = t \\ n^{\cdot,t}_{y^V=1,-ij} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} \neq t \end{cases}$$

Le scomposizioni $n^{h,t}_{y^H=1}$ e $n^{\cdot,t}_{y^H=1}$ sono ottenute seguendo lo stesso procedimento utilizzato per $n^{v,t}_{y^V=1}$ e $n^{\cdot,t}_{y^V=1}$:

$$n^{h,t}_{y^H=1} = \begin{cases} n^{h,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}\mathbb{1}_{z^H_{ijl}=t} & \text{se } x_{ij} = 1 \\ n^{h,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} = t \\ n^{h,t}_{y^H=1,-ij} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} \neq t \end{cases}$$

$$n^{\cdot,t}_{y^H=1} = \begin{cases} n^{\cdot,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1}\mathbb{1}_{z^H_{ijl}=t} & \text{se } x_{ij} = 1 \\ n^{\cdot,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} = t \\ n^{\cdot,t}_{y^H=1,-ij} & \text{se } x_{ij} = 0 \text{ e } z^*_{ij} \neq t \end{cases}$$

La probabilità di $x_{ij} = 1$ date tutte le altre variabili è quindi data da:

$$\begin{aligned} & p(x_{ij} = 1|\dots) \\ \propto & \frac{\Gamma(b^T_1 + n^1_{x_i,-j} + 1)\Gamma(b^T_1 + n^0_{x_i,-j})}{\Gamma(b^T_1 + b^T_2 + D_i)} \\ & \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V}\Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}\mathbb{1}_{z^V_{ijn}=t})}{\Gamma(\sum_{v=1}^{V}\beta^V_v + n^{\cdot,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1}\mathbb{1}_{z^V_{ijn}=t})} \end{aligned}$$

$$
\begin{aligned}
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n^{h,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n^{\cdot,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t})} \\
=&\frac{(b_1^T + n^1_{x_i,-j})\Gamma(b_1^T + n^1_{x_i,-j})\Gamma(b_2^T + n^0_{x_i,-j})}{(b_1^T + b_2^T + D_i - 1)\Gamma(b_1^T + b_2^T + D_i - 1)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} (\beta_v^V + n^{v,t}_{y^V=1,-ij} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} (\beta_v^V + n^{\cdot,t}_{y^V=1,-ij} + q)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} \Gamma(\beta_v^V + n^{v,t}_{y^V=1,-ij})}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} \Gamma(\beta_v^V + n^{\cdot,t}_{y^V=1,-ij})} \\
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} (\beta_h^H + n^{h,t}_{y^H=1,-ij} + q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} (\beta_h^H + n^{\cdot,t}_{y^H=1,-ij} + q)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} \Gamma(\beta_h^H + n^{h,t}_{y^H=1,-ij})}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} \Gamma(\beta_h^H + n^{\cdot,t}_{y^H=1,-ij})} \\
\propto&\frac{(b_1^T + n^1_{x_i,-j})}{(b_1^T + b_2^T + D_i - 1)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} (\beta_v^V + n^{v,t}_{y^V=1,-ij} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} (\beta_v^V + n^{\cdot,t}_{y^V=1,-ij} + q)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} (\beta_h^H + n^{h,t}_{y^H=1,-ij} + q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} (\beta_h^H + n^{\cdot,t}_{y^H=1,-ij} + q)}
\end{aligned}
$$

La probabilità di $x_{ij} = 0$ date tutte le altre variabili è quindi data da:

$$
\begin{aligned}
&p(x_{ij} = 0 | z^*_{ij} = t^*, \dots) \\
\propto&\frac{\Gamma(b_1^T + n^1_{x_i,-j})\Gamma(b_1^T + n^0_{x_i,-j} + 1)}{\Gamma(b_1^T + b_2^T + D_i)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^{v,t}_{y^V=1,-ij} + \mathbb{1}_{z_{ij}=t} \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot,t}_{y^V=1,-ij} + \mathbb{1}_{z_{ij}=t} \sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1})}
\end{aligned}
$$

$$
\begin{aligned}
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n_{y^H=1,-ij}^{h,t} + \mathbb{1}_{z_{ij}=t} \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n_{y^H=1,-ij}^{\cdot,t} + \mathbb{1}_{z_{ij}=t} \sum_{l=1}^{L_{ij}} \mathbb{1}_{y_{ijl}^H=1})} \\
=&\frac{\Gamma(b_1^T + n_{x_i,-j}^1)(b_2^T + n_{x_i,-j}^0)\Gamma(b_2^T + n_{x_i,-j}^0)}{(b_1^T + b_2^T + D_i - 1)\Gamma(b_1^T + b_2^T + D_i - 1)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y_{ijn}^V=1} - 1} (\beta_v^V + n_{y^V=1,-ij}^{v,t} + q)}{\prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{n=1}^{N_{ij}} \mathbb{1}_{y_{ijn}^V=1} - 1} (\sum_{v=1}^{V} \beta_v^V + n_{y^V=1,-ij}^{\cdot,t} + q)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y_{ijn}^V=1} - 1} \Gamma(\beta_v^V + n_{y^V=1,-ij}^{v,t})}{\prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{n=1}^{N_{ij}} \mathbb{1}_{y_{ijn}^V=1} - 1} \Gamma(\sum_{v=1}^{V} \beta_v^V + n_{y^V=1,-ij}^{\cdot,t})} \\
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1} - 1} (\beta_h^H + n_{y^H=1,-ij}^{h,t^*} + q)}{\prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{l=1}^{L_{ij}} \mathbb{1}_{y_{ijl}^H=1} - 1} (\sum_{h=1}^{H} \beta_h^H + n_{y^H=1,-ij}^{\cdot,t} + q)} \\
&\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1} - 1} \Gamma(\beta_h^H + n_{y^H=1,-ij}^{h,t})}{\prod_{q=0}^{\mathbb{1}_{z_{ij}=t} \sum_{l=1}^{L_{ij}} \mathbb{1}_{y_{ijl}^H=1} - 1} \Gamma(\sum_{h=1}^{H} \beta_h^H + n_{y^H=1,-ij}^{\cdot,t})} \\
\propto&\frac{(b_2^T + n_{x_i,-j}^0)}{(b_1^T + b_2^T + D_i - 1)} \\
&\times \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y_{ijn}^V=1} - 1} (\beta_v^V + n_{y^V=1,-ij}^{v,t^*} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y_{ijn}^V=1} - 1} (\sum_{v=1}^{V} \beta_v^V + n_{y^V=1,-ij}^{\cdot,t^*} + q)} \\
&\times \frac{\prod_{h=1}^{H} \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1} - 1} (\beta_h^H + n_{y^H=1,-ij}^{h,t^*} + q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y_{ijl}^H=1} - 1} (\sum_{h=1}^{H} \beta_h^H + n_{y^H=1,-ij}^{\cdot,t^*} + q)}
\end{aligned}
$$

Distribuzione del topic principale

La distribuzione del topic principale z_{ij}^* date tutte le altre variabili è proporzionale al prodotto dei blocchi (5), (8) e (10):

$$p(z_{ij}^* = t^* | \dots)$$

$$
\begin{aligned}
&\propto \prod_{u=1}^{U} \frac{\prod_{t=1}^{T} \Gamma(\alpha_t^* + n_{z_u^*}^t)}{\Gamma(\sum_{t=1}^{T} \alpha_t^* + D_u)} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n_{y^V=1}^{v,t})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n_{y^V=1}^{\cdot,t})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n_{y^H=1}^{h,t})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n_{y^H=1}^{\cdot,t})} \\
&\propto \frac{\prod_{t=1}^{T} \Gamma(\alpha_t^* + n_{z_i^*}^t)}{\Gamma(\sum_{t=1}^{T} \alpha_t^* + D_i)} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n_{y^V=1}^{v,t})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n_{y^V=1}^{\cdot,t})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n_{y^H=1}^{h,t})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n_{y^H=1}^{\cdot,t})}
\end{aligned}
$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare z_{ij}^*:

$$
\begin{aligned}
n_{z_i^*}^t &= \sum_{d=1}^{D_i} \mathbb{1}_{z_{id}^*=t} = \sum_{d \neq j} \mathbb{1}_{z_{id}^*=t} + \mathbb{1}_{z_{ij}^*=t} \\
&= n_{z_i^*,-j}^t + \mathbb{1}_{z_{ij}^*=t} = \begin{cases} n_{z_i^*,-j}^t + 1 & \text{se } z_{ij}^* = t \\ n_{z_i^*,-j}^t & \text{se } z_{ij}^* \neq t \end{cases}
\end{aligned}
$$

Le scomposizioni di $n_{y^V=1}^{v,t}$, $n_{y^V=1}^{\cdot,t}$, $n_{y^H=1}^{h,t}$ e $n_{y^H=1}^{\cdot,t}$ coincidono con quelle ottenute nella sezione precedente; per comodità, si riportano nuovamente i risultati:

$$
n_{y^V=1}^{v,t} = \begin{cases} n_{y^V=1,-ij}^{v,t} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y_{ijn}^V=1} \mathbb{1}_{z_{ijn}^V=t} & \text{se } x_{ij} = 1 \\ n_{y^V=1,-ij}^{v,t} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y_{ijn}^V=1} & \text{se } x_{ij} = 0 \text{ e } z_{ij}^* = t \\ n_{y^V=1,-ij}^{v,t} & \text{se } x_{ij} = 0 \text{ e } z_{ij}^* \neq t \end{cases}
$$

$$
n_{y^V=1}^{\cdot,t} = \begin{cases} n_{y^V=1,-ij}^{\cdot,t} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y_{ijn}^V=1} \mathbb{1}_{z_{ijn}^V=t} & \text{se } x_{ij} = 1 \\ n_{y^V=1,-ij}^{\cdot,t} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y_{ijn}^V=1} & \text{se } x_{ij} = 0 \text{ e } z_{ij}^* = t \\ n_{y^V=1,-ij}^{\cdot,t} & \text{se } x_{ij} = 0 \text{ e } z_{ij}^* \neq t \end{cases}
$$

$$
n_{y^H=1}^{h,t} = \begin{cases} n_{y^H=1,-ij}^{h,t} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1} \mathbb{1}_{z_{ijl}^H=t} & \text{se } x_{ij} = 1 \\ n_{y^H=1,-ij}^{h,t} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1} & \text{se } x_{ij} = 0 \text{ e } z_{ij}^* = t \\ n_{y^H=1,-ij}^{h,t} & \text{se } x_{ij} = 0 \text{ e } z_{ij}^* \neq t \end{cases}
$$

$$n^{\cdot,t}_{y^H=1} = \begin{cases} n^{\cdot,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1}\mathbb{1}_{z^H_{ijl}=t} & \text{se } x_{ij}=1 \\ n^{\cdot,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} & \text{se } x_{ij}=0 \text{ e } z^*_{ij}=t \\ n^{\cdot,t}_{y^H=1,-ij} & \text{se } x_{ij}=0 \text{ e } z^*_{ij}\neq t \end{cases}$$

Quindi, è possibile distinguere due casi $x_{ij}=1$ e $x_{ij}=0$; la probabilità di $z^*_{ij}=t^*$ dato $x_{ij}=1$ e tutte le altre variabili è data da:

$$\begin{aligned} & p(z^*_{ij}=t^*|x_{ij}=1,\dots) \\ \propto & \frac{\prod_{t=1}^T \Gamma(\alpha^*_t + n^t_{z^*_i,-ij} + \mathbb{1}_{z^*_{ij}=t})}{\Gamma(\sum_{t=1}^T \alpha^*_t + D_i)} \\ & \times \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}\mathbb{1}_{z^V_{ijn}=t})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ij} + \sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1}\mathbb{1}_{z^V_{ijn}=t})} \\ & \times \prod_{t=1}^T \frac{\prod_{h=1}^H \Gamma(\beta^H_h + n^{h,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}\mathbb{1}_{z^H_{ijl}=t})}{\Gamma(\sum_{h=1}^H \beta^H_h + n^{\cdot,t}_{y^H=1,-ij} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1}\mathbb{1}_{z^H_{ijl}=t})} \end{aligned}$$

Il secondo e terzo blocco, che derivano da (8) e (10), possono essere trascurati poiché non dipendono da z^*_{ij}:

$$\begin{aligned} & \propto \frac{\Gamma(\alpha^*_{t^*} + n^{t^*}_{z^*_i,-ij} + 1)\prod_{t\neq t^*}\Gamma(\alpha^*_t + n^t_{z^*_i,-ij})}{\Gamma(\sum_{t=1}^T \alpha^*_t + D_i)} \\ & \propto \frac{(\alpha^*_{t^*} + n^{t^*}_{z^*_i,-ij})\prod_{t=1}^T \Gamma(\alpha^*_t + n^t_{z^*_i,-ij})}{(\sum_{t=1}^T \alpha^*_t + D_i - 1)\Gamma(\sum_{t=1}^T \alpha^*_t + D_i - 1)} \\ & \propto \frac{\alpha_{t^*} + n^{t^*}_{z^*_i,-ij}}{\sum_{t=1}^T \alpha^*_t + D_i - 1} \end{aligned}$$

La probabilità di $z^*_{ij}=t^*$ dato $x_{ij}=0$ e tutte le altre variabili è data da:

$$\begin{aligned} & p(z^*_{ij}=t^*|x_{ij}=0,\dots) \\ \propto & \frac{\prod_{t=1}^T \Gamma(\alpha^*_t + n^t_{z^*_i,-ij} + \mathbb{1}_{z^*_{ij}=t})}{\Gamma(\sum_{t=1}^T \alpha^*_t + D_i)} \\ & \times \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ij} + \mathbb{1}_{z^*_{ij}=t}\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ij} + \mathbb{1}_{z^*_{ij}=t}\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1})} \\ & \times \prod_{t=1}^T \frac{\prod_{h=1}^H \Gamma(\beta^H_h + n^{h,t}_{y^H=1,-ij} + \mathbb{1}_{z^*_{ij}=t}\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1})}{\Gamma(\sum_{h=1}^H \beta^H_h + n^{\cdot,t}_{y^H=1,-ij} + \mathbb{1}_{z^*_{ij}=t}\sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1})} \end{aligned}$$

$$
\begin{aligned}
=&\frac{\Gamma(\alpha^*_{t^*}+n^{t^*}_{z_i^*,-ij}+1)\prod_{t\neq t^*}\Gamma(\alpha^*_t+n^{t}_{z_i^*,-ij})}{\Gamma(\sum_{t=1}^T\alpha^*_t+D_i)}\\
&\times\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t^*}_{y^V=1,-ij}+\sum_{n=1}^{N_{ij}}\mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^*}_{y^V=1,-ij}+\sum_{n=1}^{N_{ij}}\mathbb{1}_{y^V_{ijn}=1})}\\
&\times\prod_{t\neq t^*}\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ij})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ij})}\\
&\times\frac{\prod_{h=1}^H\Gamma(\beta^H_h+n^{h,t^*}_{y^H=1,-ij}+\sum_{l=1}^{L_{ij}}\mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1})}{\Gamma(\sum_{h=1}^H\beta^H_h+n^{\cdot,t^*}_{y^H=1,-ij}+\sum_{l=1}^{L_{ij}}\mathbb{1}_{y^H_{ijl}=1})}\\
&\times\prod_{t\neq t^*}\frac{\prod_{h=1}^H\Gamma(\beta^H_h+n^{h,t}_{y^H=1,-ij})}{\Gamma(\sum_{h=1}^H\beta^H_h+n^{\cdot,t}_{y^H=1,-ij})}\\
=&\frac{(\alpha^*_{t^*}+n^{t^*}_{z_i^*,-ij})\Gamma(\alpha^*_{t^*}+n^{t^*}_{z_i^*,-ij})\prod_{t\neq t^*}\Gamma(\alpha^*_t+n^{t}_{z_i^*,-ij})}{(\sum_{t=1}^T\alpha^*_t+D_i-1)\Gamma(\sum_{t=1}^T\alpha^*_t+D_i-1)}\\
&\times\frac{\prod_{v=1}^V\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}-1}(\beta^V_v+n^{v,t^*}_{y^V=1,-ij}+q)\Gamma(\beta^V_v+n^{v,t^*}_{y^V=1,-ij})}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{y^V_{ijn}=1}-1}(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^*}_{y^V=1,-ij}+q)\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^*}_{y^V=1,-ij})}\\
&\times\prod_{t\neq t^*}\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ij})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ij})}\\
&\times\frac{\prod_{h=1}^H\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}-1}(\beta^H_h+n^{h,t^*}_{y^H=1,-ij}+q)\Gamma(\beta^H_h+n^{h,t^*}_{y^H=1,-ij})}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{y^H_{ijl}=1}-1}(\sum_{h=1}^H\beta^H_h+n^{\cdot,t^*}_{y^H=1,-ij}+q)\Gamma(\sum_{h=1}^H\beta^H_h+n^{\cdot,t^*}_{y^H=1,-ij})}\\
&\times\prod_{t\neq t^*}\frac{\prod_{h=1}^H\Gamma(\beta^H_h+n^{h,t}_{y^H=1,-ij})}{\Gamma(\sum_{h=1}^H\beta^H_h+n^{\cdot,t}_{y^H=1,-ij})}\\
=&\frac{\alpha^*_{t^*}+n^{t^*}_{z_i^*,-ij}}{\sum_{t=1}^T\alpha^*_t+D_i-1}\frac{\prod_{t=1}^T\Gamma(\alpha^*_t+n^{t}_{z_i^*,-ij})}{\Gamma(\sum_{t=1}^T\alpha^*_t+D_i)}\\
&\times\frac{\prod_{v=1}^V\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}-1}(\beta^V_v+n^{v,t^*}_{y^V=1,-ij}+q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{y^V_{ijn}=1}-1}(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^*}_{y^V=1,-ij}+q)}\\
&\times\prod_{t=1}^T\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ij})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ij})}
\end{aligned}
$$

$$
\begin{aligned}
&\times \frac{\prod_{h=1}^{H}\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}-1}(\beta_h^H + n^{h,t^*}_{y^H=1,-ij}+q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{y^H_{ijl}=1}-1}(\sum_{h=1}^{H}\beta_h^H + n^{\cdot,t^*}_{y^H=1,-ij}+q)} \\
&\times \prod_{t=1}^{T}\frac{\prod_{h=1}^{H}\Gamma(\beta_h^H + n^{h,t}_{y^H=1,-ij})}{\Gamma(\sum_{h=1}^{H}\beta_h^H + n^{\cdot,t}_{y^H=1,-ij})} \\
\propto &\frac{\alpha^*_{t^*} + n^{t^*}_{z_i^*,-ij}}{\sum_{t=1}^{T}\alpha_t^* + D_i - 1} \\
&\times \frac{\prod_{v=1}^{V}\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}-1}(\beta_v^V + n^{v,t^*}_{y^V=1,-ij}+q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{y^V_{ijn}=1}-1}(\sum_{v=1}^{V}\beta_v^V + n^{\cdot,t^*}_{y^V=1,-ij}+q)} \\
&\times \frac{\prod_{h=1}^{H}\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}-1}(\beta_h^H + n^{h,t^*}_{y^H=1,-ij}+q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{y^H_{ijl}=1}-1}(\sum_{h=1}^{H}\beta_h^H + n^{\cdot,t^*}_{y^H=1,-ij}+q)}
\end{aligned}
$$

Distribuzione della presenza di un topic in un documento

La distribuzione di $\lambda_{ij,k}$ date tutte le altre variabili è proporzionale al prodotto dei blocchi (1) e (6):

$$
\begin{aligned}
&p(\lambda_{ij,k}|\dots) \\
\propto &\frac{\Gamma(b_1^\delta + n_\lambda^1)\Gamma(b_2^\delta + n_\lambda^0)}{\Gamma(b_1^\delta + b_2^\delta + DT)} \\
&\times \prod_{u=1}^{U}\prod_{d=1}^{D_u}\frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ud,t}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud},z^H_{ud}})}{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ud,t}\alpha_t + N_{ud} + L_{ud})} \\
\propto &\frac{\Gamma(b_1^\delta + n_\lambda^1)\Gamma(b_2^\delta + n_\lambda^0)}{\Gamma(b_1^\delta + b_2^\delta + DT)} \\
&\times \frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t)}{\Gamma(\alpha_0 + \lambda_{ij,k}\alpha_k)}\frac{\Gamma(\alpha_0 + \lambda_{ij,k}\alpha_k + n^k_{z^V_{ij},z^H_{ij}})}{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij})} \\
\propto &\frac{\Gamma(b_1^\delta + n_\lambda^1)\Gamma(b_2^\delta + n_\lambda^0)}{\Gamma(b_1^\delta + b_2^\delta + DT)} \\
&\times \frac{\Gamma(\alpha_0 + \lambda_{ij,k}\alpha_k + n^k_{z^V_{ij},z^H_{ij}})}{\Gamma(\alpha_0 + \lambda_{ij,k}\alpha_k)}\frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t)}{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij})} \\
\propto &\frac{\Gamma(b_1^\delta + n_\lambda^1)\Gamma(b_2^\delta + n_\lambda^0)}{\Gamma(b_1^\delta + b_2^\delta + DT)}
\end{aligned}
$$

$$\times \frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+\lambda_{ij,k}\alpha_k+q)\Gamma(\alpha_0+\lambda_{ij,k}\alpha_k)}{\Gamma(\alpha_0+\lambda_{ij,k}\alpha_k)}$$
$$\times \frac{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+q)\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t)}$$
$$\propto \frac{\Gamma(b_1^\delta+n_\lambda^1)\Gamma(b_2^\delta+n_\lambda^0)}{\Gamma(b_1^\delta+b_2^\delta+DT)} \times \frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+\lambda_{ij,k}\alpha_k+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+q)}$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare $\lambda_{ij,k}$:

$$\begin{aligned}
n_\lambda^1 &= \sum_{udt}\mathbb{1}_{\lambda_{ud,t}=1} = \sum_{udt\neq ijk}\mathbb{1}_{\lambda_{ud,t}=1}+\mathbb{1}_{\lambda_{ij,k}=1} \\
&= n^1_{\lambda,-ijk}+\mathbb{1}_{\lambda_{ij,k}=1} = \begin{cases} n^1_{\lambda,-ijk}+1 & \text{se } \lambda_{ij,k}=1 \\ n^1_{\lambda,-ijk} & \text{se } \lambda_{ij,k}=0 \end{cases} \\
n_\lambda^0 &= \sum_{udt}\mathbb{1}_{\lambda_{ud,t}=0} = \sum_{udt\neq ijk}\mathbb{1}_{\lambda_{ud,t}=0}+\mathbb{1}_{\lambda_{ij,k}=0} \\
&= n^0_{\lambda,-ijk}+\mathbb{1}_{\lambda_{ij,k}=0} = \begin{cases} n^0_{\lambda,-ijk} & \text{se } \lambda_{ij,k}=1 \\ n^0_{\lambda,-ijk}+1 & \text{se } \lambda_{ij,k}=0 \end{cases}
\end{aligned}$$

Inoltre vale la seguente scomposizione della somma dei paramatri corrispondenti ai topic attivi del documento:

$$\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t = \sum_{t\neq k}\lambda_{ij,t}\alpha_t+\lambda_{ij,k}\alpha_k = \begin{cases} \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t & \text{se } \lambda_{ij,k}=1 \\ \sum_{t\neq k}\lambda_{ij,t}\alpha_t & \text{se } \lambda_{ij,k}=0 \end{cases}$$

La probabilità di $\lambda_{ij,k}=1$ date tutte le altre variabili è quindi data da:

$$\begin{aligned}
&p(\lambda_{ij,k}=1|\dots) \\
&\propto \frac{\Gamma(b_1^\delta+n^1_{\lambda,-j}+1)\Gamma(b_2^\delta+n^0_{\lambda,-j})}{\Gamma(b_1^\delta+b_2^\delta+DT)} \\
&\quad\times \frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+\alpha_k+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+q)} \\
&\propto \frac{(b_1^\delta+n^1_{\lambda,-j})\Gamma(b_1^\delta+n^1_{\lambda,-j})\Gamma(b_2^\delta+n^0_{\lambda,-j})}{(b_1^\delta+b_2^\delta+DT-1)\Gamma(b_1^\delta+b_2^\delta+DT-1)}
\end{aligned}$$

$$\times \frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+\alpha_k+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+q)}$$
$$\propto \frac{b_1^\delta+n^1_{\lambda,-j}}{b_1^\delta+b_2^\delta+DT-1}\times\frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+\alpha_k+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+q)}$$

La probabilità di $\lambda_{ij,k}=0$ date tutte le altre variabili è quindi data da:

$$p(\lambda_{ij,k}=0|\dots)$$
$$\propto\frac{\Gamma(b_1^\delta+n^1_{\lambda,-j})\Gamma(b_2^\delta+n^0_{\lambda,-j}+1)}{\Gamma(b_1^\delta+b_2^\delta+DT)}$$
$$\times\frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t\neq k}\lambda_{ij,t}\alpha_t+q)}$$
$$\propto\frac{\Gamma(b_1^\delta+n^1_{\lambda,-j})(b_2^\delta+n^0_{\lambda,-j})\Gamma(b_2^\delta+n^0_{\lambda,-j})}{(b_1^\delta+b_2^\delta+DT-1)\Gamma(b_1^\delta+b_2^\delta+DT-1)}$$
$$\times\frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t\neq k}\lambda_{ij,t}\alpha_t+q)}$$
$$\propto\frac{b_2^\delta+n^0_{\lambda,-j}}{b_1^\delta+b_2^\delta+DT-1}\times\frac{\prod_{q=0}^{n^k_{z^V_{ij},z^H_{ij}}-1}(\alpha_0+q)}{\prod_{q=0}^{N_{ij}+L_{ij}-1}(T\alpha_0+\sum_{t\neq k}^{T}\lambda_{ij,t}\alpha_t+q)}$$

Distribuzione dell'origine di una parola

La distribuzione di y^V_{ijk} date tutte le altre variabili è proporzionale al prodotto dei blocchi (2), (7) e (8):

$$p(y^V_{ijk}|\dots)$$
$$\propto\frac{\Gamma(b_1^V+n^1_{y^V})\Gamma(b_2^V+n^0_{y^V})}{\Gamma(b_1^V+b_2^V+N)}\times\frac{\prod_{v=1}^{V}\Gamma(\beta_v^V+n^v_{y^V=0})}{\Gamma(\sum_{v=1}^{V}\beta_v^V+n^{\cdot}_{y^V=0})}$$
$$\times\prod_{t=1}^{T}\frac{\prod_{v=1}^{V}\Gamma(\beta_v^V+n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V}\beta_v^V+n^{\cdot,t}_{y^V=1})}$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare y^V_{ijk}:

$$
\begin{aligned}
n^1_{y^V} &= \sum_{udn} \mathbb{1}_{y^V_{udn}=1} = \sum_{udn\neq ijk} \mathbb{1}_{y^V_{udn}=1} + \mathbb{1}_{y^V_{ijk}=1} \\
&= n^1_{y^V,-ijk} + \mathbb{1}_{y^V_{ijk}=1} = \begin{cases} n^1_{y^V,-ijk}+1 & \text{se } y^V_{ijk}=1 \\ n^1_{y^V,-ijk} & \text{se } y^V_{ijk}=0 \end{cases} \\
n^0_{y^V} &= \sum_{udn} \mathbb{1}_{y^V_{udn}=0} = \sum_{udn\neq ijk} \mathbb{1}_{y^V_{udn}=0} + \mathbb{1}_{y^V_{ijk}=0} \\
&= n^0_{y^V,-ijk} + \mathbb{1}_{y^V_{ijk}=0} = \begin{cases} n^0_{y^V,-ijk} & \text{se } y^V_{ijk}=1 \\ n^0_{y^V,-ijk}+1 & \text{se } y^V_{ijk}=0 \end{cases} \\
n^v_{y^V=0} &= \sum_{udn} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0} = \sum_{udn\neq ijk} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0} + \mathbb{1}_{w_{ijk}=v}\mathbb{1}_{y^V_{ijk}=0} \\
&= n^v_{y^V=0,-ijk} + \mathbb{1}_{w_{ijk}=v}\mathbb{1}_{y^V_{ijk}=0} = \begin{cases} n^v_{y^V=0,-ijk}+1 & \text{se } w_{ijk}=v \text{ e } y^V_{ijk}=0 \\ n^v_{y^V=0,-ijk} & \text{altrimenti} \end{cases} \\
n^{\cdot}_{y^V=0} &= \sum_{udn} \mathbb{1}_{y^V_{udn}=0} = \sum_{udn\neq ijk} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0} + \mathbb{1}_{y^V_{ijk}=0} \\
&= n^{\cdot}_{y^V=0,-ijk} + \mathbb{1}_{y^V_{ijk}=0} = \begin{cases} n^{\cdot}_{y^V=0,-ijk}+1 & \text{se } y^V_{ijk}=0 \\ n^{\cdot}_{y^V=0,-ijk} & \text{altrimenti} \end{cases} \\
n^{v,t}_{y^V=1} &= \sum_{u=1}^{U}\sum_{d=1}^{D_u}\sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t}\mathbb{1}_{x_{ud}=0}) \\
&= n^{v,t}_{y^V=1,-ijk} + \mathbb{1}_{w_{ijk}=v}\mathbb{1}_{y^V_{ijk}=1}(\mathbb{1}_{z^V_{ijk}=t}\mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0}) \\
&= \begin{cases} n^{v,t}_{y^V=1,-ijk}+1 & \text{se } w_{ijk}=v,\ y^V_{ijk}=1,\ x_{ij}=1 \text{ e } z^V_{ijk}=t \\ n^{v,t}_{y^V=1,-ijk}+1 & \text{se } w_{ijk}=v,\ y^V_{ijk}=1,\ x_{ij}=0 \text{ e } z^*_{ij}=t \\ n^{v,t}_{y^V=1,-ijk} & \text{altrimenti} \end{cases} \\
n^{\cdot,t}_{y^V=1} &= \sum_{v=1}^{V} n^{v,t}_{y^V=1} = n^{\cdot,t}_{y^V=1,-ijk} + \mathbb{1}_{y^V_{ijk}=1}(\mathbb{1}_{z^V_{ijk}=t}\mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0}) \\
&= \begin{cases} n^{\cdot,t}_{y^V=1,-ijk}+1 & \text{se } y^V_{ijk}=1,\ x_{ij}=1 \text{ e } z^V_{ijk}=t \\ n^{\cdot,t}_{y^V=1,-ijk}+1 & \text{se } y^V_{ijk}=1,\ x_{ij}=0 \text{ e } z^*_{ij}=t \\ n^{\cdot,t}_{y^V=1,-ijk} & \text{altrimenti} \end{cases}
\end{aligned}
$$

Quindi, è possibile distinguere due casi, $x_{ij} = 1$ e $x_{ij} = 0$, quando si calcola la probabilità di $y^V_{ijk} = 1$ date tutte le altre variabili; la prima è data da:

$$
\begin{aligned}
&p(y^V_{ijk} = 1 | x_{ij} = 1, z^V_{ijk} = t^V, w_{ijk} = w^*, \dots) \\
\propto &\frac{\Gamma(b^V_1 + n^1_{y^V,-ijk} + 1)\Gamma(b^V_2 + n^0_{y^V,-ijk})}{\Gamma(b^V_1 + b^V_2 + N)} \times \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^v_{y^V=0,-ijk})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot}_{y^V=0,-ijk})} \\
&\times \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ijk} + \mathbb{1}_{w_{ijk}=v}\mathbb{1}_{z^V_{ijk}=t})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ijk} + \mathbb{1}_{z^V_{ijk}=t})} \\
\propto &\frac{(b^V_1 + n^1_{y^V,-ijk})\Gamma(b^V_1 + n^1_{y^V,-ijk})\Gamma(b^V_2 + n^0_{y^V,-ijk})}{(b^V_1 + b^V_2 + N - 1)\Gamma(b^V_1 + b^V_2 + N - 1)} \\
&\times \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t^V}_{y^V=1,-ijk} + \mathbb{1}_{w_{ijk}=v})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t^V}_{y^V=1,-ijk} + 1)} \prod_{t \neq t^V} \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ijk})} \\
\propto &\frac{b^V_1 + n^1_{y^V,-ijk}}{b^V_1 + b^V_2 + N - 1} \frac{\Gamma(b^V_1 + n^1_{y^V,-ijk})\Gamma(b^V_2 + n^0_{y^V,-ijk})}{\Gamma(b^V_1 + b^V_2 + N - 1)} \\
&\times \frac{\Gamma(\beta^V_{w^*} + n^{w^*,t^V}_{y^V=1,-ijk} + 1)\prod_{v \neq w^*} \Gamma(\beta^V_v + n^{v,t^V}_{y^V=1,-ijk})}{(\sum_{v=1}^V \beta^V_v + n^{\cdot,t^V}_{y^V=1,-ijk})\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t^V}_{y^V=1,-ijk})} \\
&\times \prod_{t \neq t^V} \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ijk})} \\
\propto &\frac{b^V_1 + n^1_{y^V,-ijk}}{b^V_1 + b^V_2 + N - 1} \frac{\Gamma(b^V_1 + n^1_{y^V,-ijk})\Gamma(b^V_2 + n^0_{y^V,-ijk})}{\Gamma(b^V_1 + b^V_2 + N - 1)} \\
&\times \frac{\beta^V_{w^*} + n^{w^*,t^V}_{y^V=1,-ijk}}{\sum_{v=1}^V \beta^V_v + n^{\cdot,t^V}_{y^V=1,-ijk}} \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ijk})} \\
\propto &\frac{b^V_1 + n^1_{y^V,-ijk}}{b^V_1 + b^V_2 + N - 1} \times \frac{\beta^V_{w^*} + n^{w^*,t^V}_{y^V=1,-ijk}}{\sum_{v=1}^V \beta^V_v + n^{\cdot,t^V}_{y^V=1,-ijk}}
\end{aligned}
$$

La seconda è data da:

$$
\begin{aligned}
&p(y^V_{ijk} = 1 | x_{ij} = 0, z^*_{ij} = t^*, w_{ijk} = w^*, \dots) \\
\propto &\frac{\Gamma(b^V_1 + n^1_{y^V,-ijk} + 1)\Gamma(b^V_2 + n^0_{y^V,-ijk})}{\Gamma(b^V_1 + b^V_2 + N)} \times \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^v_{y^V=0,-ijk})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot}_{y^V=0,-ijk})} \\
&\times \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta^V_v + n^{v,t}_{y^V=1,-ijk} + \mathbb{1}_{w_{ijk}=v}\mathbb{1}_{z^*_{ij}=t})}{\Gamma(\sum_{v=1}^V \beta^V_v + n^{\cdot,t}_{y^V=1,-ijk} + \mathbb{1}_{z^*_{ij}=t})}
\end{aligned}
$$

$$
\begin{aligned}
&\propto \frac{(b_1^V + n^1_{y^V,-ijk})\Gamma(b_1^V + n^1_{y^V,-ijk})\Gamma(b_2^V + n^0_{y^V,-ijk})}{(b_1^V + b_2^V + N - 1)\Gamma(b_1^V + b_2^V + N - 1)} \\
&\quad \times \frac{\prod_{v=1}^V \Gamma(\beta_v^V + n^{v,t^*}_{y^V=1,-ijk} + \mathbb{1}_{w_{ijk}=v})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot,t^*}_{y^V=1,-ijk} + 1)} \prod_{t\neq t^*} \frac{\prod_{v=1}^V \Gamma(\beta_v^V + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot,t}_{y^V=1,-ijk})} \\
&\propto \frac{b_1^V + n^1_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \frac{\Gamma(b_1^V + n^1_{y^V,-ijk})\Gamma(b_2^V + n^0_{y^V,-ijk})}{\Gamma(b_1^V + b_2^V + N - 1)} \\
&\quad \times \frac{\Gamma(\beta^V_{w^*} + n^{w^*,t^*}_{y^V=1,-ijk} + 1)\prod_{v\neq w^*}\Gamma(\beta_v^V + n^{v,t^*}_{y^V=1,-ijk})}{(\sum_{v=1}^V \beta_v^V + n^{\cdot,t^*}_{y^V=1,-ijk})\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot,t^*}_{y^V=1,-ijk})} \\
&\quad \times \prod_{t\neq t^*} \frac{\prod_{v=1}^V \Gamma(\beta_v^V + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot,t}_{y^V=1,-ijk})} \\
&\propto \frac{b_1^V + n^1_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \frac{\Gamma(b_1^V + n^1_{y^V,-ijk})\Gamma(b_2^V + n^0_{y^V,-ijk})}{\Gamma(b_1^V + b_2^V + N - 1)} \\
&\quad \times \frac{\beta^V_{w^*} + n^{w^*,t^*}_{y^V=1,-ijk}}{\sum_{v=1}^V \beta_v^V + n^{\cdot,t^*}_{y^V=1,-ijk}} \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta_v^V + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot,t}_{y^V=1,-ijk})} \\
&\propto \frac{b_1^V + n^1_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \times \frac{\beta^V_{w^*} + n^{w^*,t^*}_{y^V=1,-ijk}}{\sum_{v=1}^V \beta_v^V + n^{\cdot,t^*}_{y^V=1,-ijk}}
\end{aligned}
$$

La probabilità di $y^V_{ijk} = 0$ date tutte le altre variabili è data da:

$$
\begin{aligned}
&p(y^V_{ijk} = 0 | w_{ijk} = w^*, \ldots) \\
&\propto \frac{\Gamma(b_1^V + n^1_{y^V,-ijk})\Gamma(b_2^V + n^0_{y^V,-ijk} + 1)}{\Gamma(b_1^V + b_2^V + N)} \times \frac{\prod_{v=1}^V \Gamma(\beta_v^V + n^v_{y^V=0,-ijk} + \mathbb{1}_{w_{ijk}=v})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot}_{y^V=0,-ijk} + 1)} \\
&\quad \times \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta_v^V + n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot,t}_{y^V=1,-ijk})} \\
&\propto \frac{\Gamma(b_1^V + n^1_{y^V,-ijk})(b_2^V + n^0_{y^V,-ijk})\Gamma(b_2^V + n^0_{y^V,-ijk})}{(b_1^V + b_2^V + N - 1)\Gamma(b_1^V + b_2^V + N - 1)} \\
&\quad \times \frac{\Gamma(\beta^V_{w^*} + n^{w^*}_{y^V=0,-ijk} + 1)\prod_{v\neq w^*}\Gamma(\beta_v^V + n^v_{y^V=0,-ijk})}{(\sum_{v=1}^V \beta_v^V + n^{\cdot}_{y^V=0,-ijk})\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot}_{y^V=0,-ijk})} \\
&\propto \frac{b_2^V + n^0_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \frac{\Gamma(b_1^V + n^1_{y^V,-ijk})\Gamma(b_2^V + n^0_{y^V,-ijk})}{\Gamma(b_1^V + b_2^V + N - 1)} \\
&\quad \times \frac{\beta^V_{w^*} + n^{w^*}_{y^V=0,-ijk}}{(\sum_{v=1}^V \beta_v^V + n^{\cdot}_{y^V=0,-ijk})} \frac{\Gamma(\beta^V_{w^*} + n^{w^*}_{y^V=0,-ijk})\prod_{v\neq w^*}\Gamma(\beta_v^V + n^v_{y^V=0,-ijk})}{\Gamma(\sum_{v=1}^V \beta_v^V + n^{\cdot}_{y^V=0,-ijk})}
\end{aligned}
$$

$$\propto \frac{b_2^V + n^0_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \times \frac{\beta^V_{w^*} + n^{w^*}_{y^V=0,-ijk}}{\sum_{v=1}^V \beta_v^V + n^{\cdot}_{y^V=0,-ijk}}$$

Distribuzione dell'origine di un hashtag

La distribuzione di y^H_{ijk} date tutte le altre variabili è proporzionale al prodotto dei blocchi (3), (9) e (10):

$$\begin{aligned} & p(y^H_{ijk}|\dots) \\ \propto & \frac{\Gamma(b_1^H + n^1_{y^H})\Gamma(b_2^H + n^0_{y^H})}{\Gamma(b_1^H + b_2^H + L)} \times \frac{\prod_{h=1}^H \Gamma(\beta_h^H + n^h_{y^H=0})}{\Gamma(\sum_{h=1}^H \beta_h^H + n^{\cdot}_{y^H=0})} \\ & \times \prod_{t=1}^T \frac{\prod_{h=1}^H \Gamma(\beta_h^H + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^H \beta_h^H + n^{\cdot,t}_{y^H=1})} \end{aligned}$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare y^H_{ijk}:

$$\begin{aligned} n^1_{y^H} &= \sum_{udl} \mathbb{1}_{y^H_{udl}=1} = \sum_{udl \neq ijk} \mathbb{1}_{y^H_{udl}=1} + \mathbb{1}_{y^H_{ijk}=1} \\ &= n^1_{y^H,-ijk} + \mathbb{1}_{y^H_{ijk}=1} = \begin{cases} n^1_{y^H,-ijk} + 1 & \text{se } y^H_{ijk} = 1 \\ n^1_{y^H,-ijk} & \text{se } y^H_{ijk} = 0 \end{cases} \\ n^0_{y^H} &= \sum_{udl} \mathbb{1}_{y^H_{udl}=0} = \sum_{udl \neq ijk} \mathbb{1}_{y^H_{udl}=0} + \mathbb{1}_{y^H_{ijk}=0} \\ &= n^0_{y^H,-ijk} + \mathbb{1}_{y^H_{ijk}=0} = \begin{cases} n^0_{y^H,-ijk} & \text{se } y^H_{ijk} = 1 \\ n^0_{y^H,-ijk} + 1 & \text{se } y^H_{ijk} = 0 \end{cases} \\ n^h_{y^H=0} &= \sum_{udl} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=0} = \sum_{udl \neq ijk} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=0} + \mathbb{1}_{h_{ijk}=h} \mathbb{1}_{y^H_{ijk}=0} \\ &= n^h_{y^H=0,-ijk} + \mathbb{1}_{h_{ijk}=h} \mathbb{1}_{y^H_{ijk}=0} = \begin{cases} n^h_{y^H=0,-ijk} + 1 & \text{se } h_{ijk} = h \text{ e } y^H_{ijk} = 0 \\ n^h_{y^H=0,-ijk} & \text{altrimenti} \end{cases} \\ n^{\cdot}_{y^H=0} &= \sum_{udl} \mathbb{1}_{y^H_{udl}=0} = \sum_{udl \neq ijk} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=0} + \mathbb{1}_{y^H_{ijk}=0} \\ &= n^{\cdot}_{y^H=0,-ijk} + \mathbb{1}_{y^H_{ijk}=0} = \begin{cases} n^{\cdot}_{y^H=0,-ijk} + 1 & \text{se } y^H_{ijk} = 0 \\ n^{\cdot}_{y^H=0,-ijk} & \text{altrimenti} \end{cases} \\ n^{h,t}_{y^H=1} &= \sum_{u=1}^U \sum_{d=1}^{D_u} \sum_{l=1}^{L_{ud}} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} (\mathbb{1}_{z^H_{udl}=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0}) \end{aligned}$$

$$= n^{h,t}_{y^H=1,-ijk} + \mathbb{1}_{h_{ijk}=h}\mathbb{1}_{y^H_{ijk}=1}(\mathbb{1}_{z^H_{ijk}=t}\mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0})$$

$$= \begin{cases} n^{h,t}_{y^H=1,-ijk} + 1 & \text{se } h_{ijk} = h,\ y^H_{ijk} = 1,\ x_{ij} = 1 \text{ e } z^H_{ijk} = t \\ n^{h,t}_{y^H=1,-ijk} + 1 & \text{se } h_{ijk} = h,\ y^H_{ijk} = 1,\ x_{ij} = 0 \text{ e } z^*_{ij} = t \\ n^{h,t}_{y^H=1,-ijk} & \text{altrimenti} \end{cases}$$

$$n^{\cdot,t}_{y^H=1} = \sum_{h=1}^{H} n^{h,t}_{y^H=1} = n^{\cdot,t}_{y^H=1,-ijk} + \mathbb{1}_{y^H_{ijk}=1}(\mathbb{1}_{z^H_{ijk}=t}\mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0})$$

$$= \begin{cases} n^{\cdot,t}_{y^H=1,-ijk} + 1 & \text{se } y^H_{ijk} = 1,\ x_{ij} = 1 \text{ e } z^H_{ijk} = t \\ n^{\cdot,t}_{y^H=1,-ijk} + 1 & \text{se } y^H_{ijk} = 1,\ x_{ij} = 0 \text{ e } z^*_{ij} = t \\ n^{\cdot,t}_{y^H=1,-ijk} & \text{altrimenti} \end{cases}$$

Quindi, è possibile distinguere due casi, $x_{ij} = 1$ e $x_{ij} = 0$, quando si calcola la probabilità di $y^H_{ijk} = 1$ date tutte le altre variabili; la prima è data da:

$$p(y^H_{ijk} = 1 | x_{ij} = 1, z^H_{ijk} = t^H, h_{ijk} = h^*, \ldots)$$
$$\propto \frac{b^H_1 + n^1_{y^H,-ijk}}{b^H_1 + b^H_2 + L - 1} \times \frac{\beta^H_{h^*} + n^{h^*,t^H}_{y^H=1,-ijk}}{\sum_{h=1}^{H} \beta^H_h + n^{\cdot,t^H}_{y^H=1,-ijk}}$$

La seconda è data da:

$$p(y^H_{ijk} = 1 | x_{ij} = 0, z^*_{ij} = t^*, h_{ijk} = h^*, \ldots)$$
$$\propto \frac{b^H_1 + n^1_{y^H,-ijk}}{b^H_1 + b^H_2 + L - 1} \times \frac{\beta^H_{h^*} + n^{h^*,t^*}_{y^H=1,-ijk}}{\sum_{h=1}^{H} \beta^H_h + n^{\cdot,t^*}_{y^H=1,-ijk}}$$

La probabilità di $y^H_{ijk} = 0$ date tutte le altre variabili è data da:

$$p(y^H_{ijk} = 0 | h_{ijk} = h^*, \ldots)$$
$$\propto \frac{b^H_2 + n^0_{y^H,-ijk}}{b^H_1 + b^H_2 + L - 1} \times \frac{\beta^H_{h^*} + n^{h^*}_{y^H=0,-ijk}}{\sum_{h=1}^{H} \beta^H_h + n^{\cdot}_{y^H=0,-ijk}}$$

Il procedimento è stato omesso poiché coincide con quello esposto nella sezione precedente.

Distribuzione del topic di una parola

La distribuzione di z^V_{ijk} date tutte le altre variabili è proporzionale al prodotto dei blocchi (6) e (8):

$$\begin{aligned}
&p(z^V_{ijk}|\dots)\\
&\propto \prod_{u=1}^{U}\prod_{d=1}^{D_u}\frac{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ud,t}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t+n^t_{z^V_{ud},z^H_{ud}})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ud,t}\alpha_t+N_{ud}+L_{ud})}\\
&\quad\times\prod_{t=1}^{T}\frac{\prod_{v=1}^{V}\Gamma(\beta^V_v+n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V}\beta^V_v+n^{\cdot,t}_{y^V=1})}\\
&\propto\frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^t_{z^V_{ij},z^H_{ij}})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij})}\times\prod_{t=1}^{T}\frac{\prod_{v=1}^{V}\Gamma(\beta^V_v+n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V}\beta^V_v+n^{\cdot,t}_{y^V=1})}
\end{aligned}$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare z^V_{ijk}:

$$\begin{aligned}
n^t_{z^V_{ij},z^H_{ij}}&=\sum_{n=1}^{N_{ij}}\mathbb{1}_{z^V_{ijn}=t}+\sum_{l=1}^{L_{ij}}\mathbb{1}_{z^H_{ijl}=t}=\sum_{n\neq k}\mathbb{1}_{z^V_{ijn}=t}+\sum_{l=1}^{L_{ij}}\mathbb{1}_{z^H_{ijl}=t}+\mathbb{1}_{z^V_{ijk}=t}\\
&=n^t_{z^V_{ij,-k},z^H_{ij}}+\mathbb{1}_{z^V_{ijk}=t}=\begin{cases}n^t_{z^V_{ij,-k},z^H_{ij}}+1 & \text{se } z^V_{ijk}=t\\ n^t_{z^V_{ij,-k},z^H_{ij}} & \text{se } z^V_{ijk}\neq t\end{cases}\\
n^{v,t}_{y^V=1}&=\sum_{u=1}^{U}\sum_{d=1}^{D_u}\sum_{n=1}^{N_{ud}}\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}(\mathbb{1}_{z^V_{udn}=t}\mathbb{1}_{x_{ud}=1}+\mathbb{1}_{z^*_{ud}=t}\mathbb{1}_{x_{ud}=0})\\
&=n^{v,t}_{y^V=1,-ijk}+\mathbb{1}_{w_{ijk}=v}\mathbb{1}_{y^V_{ijk}=1}(\mathbb{1}_{z^V_{ijk}=t}\mathbb{1}_{x_{ij}=1}+\mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0})\\
&=\begin{cases}n^{v,t}_{y^V=1,-ijk}+\mathbb{1}_{w_{ijk}=v}\mathbb{1}_{z^V_{ijk}=t} & \text{se } y^V_{ijk}=1,\ x_{ij}=1\\ n^{v,t}_{y^V=1,-ijk}+\mathbb{1}_{w_{ijk}=v}\mathbb{1}_{z^*_{ij}=t} & \text{se } y^V_{ijk}=1,\ x_{ij}=0\\ n^{v,t}_{y^V=1,-ijk} & \text{se } y^V_{ijk}=0\end{cases}\\
n^{\cdot,t}_{y^V=1}&=\sum_{v=1}^{V}n^{v,t}_{y^V=1}=n^{\cdot,t}_{y^V=1,-ijk}+\mathbb{1}_{y^V_{ijk}=1}(\mathbb{1}_{z^V_{ijk}=t}\mathbb{1}_{x_{ij}=1}+\mathbb{1}_{z^*_{ij}=t}\mathbb{1}_{x_{ij}=0})\\
&=\begin{cases}n^{\cdot,t}_{y^V=1,-ijk}+\mathbb{1}_{z^V_{ijk}=t} & \text{se } y^V_{ijk}=1,\ x_{ij}=1\\ n^{\cdot,t}_{y^V=1,-ijk}+\mathbb{1}_{z^*_{ij}=t} & \text{se } y^V_{ijk}=1,\ x_{ij}=0\\ n^{\cdot,t}_{y^V=1,-ijk} & \text{se } y^V_{ijk}=0\end{cases}
\end{aligned}$$

I conteggi $n^{v,t}_{y^V=1}$ e $n^{\cdot,t}_{y^V=1}$ dipendono da z^V_{ijk} solo se $y^V_{ijk}=1$ e $x_{ij}=1$, è quindi possibile distinguere tre casi, $(y^V_{ijk},x_{ij})=(1,1)$, $(y^V_{ijk},x_{ij})=(1,0)$ e $y^V_{ijk}=0$, quando si calcola la probabilità di $z^V_{ijk}=t^V$ date tutte le altre variabili; la prima è data da:

$$
\begin{aligned}
&p(z^V_{ijk}=t^V|y^V_{ijk}=1,x_{ij}=1,z^V_{ijk}=t^V,w_{ijk}=w^*,\ldots)\\
\propto&\frac{\prod_{t=1}^T\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^t_{z^V_{ij,-k},z^H_{ij}}+\mathbb{1}_{z^V_{ijk}=t})}{\Gamma(T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij})}\\
&\times\prod_{t=1}^T\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ijk}+\mathbb{1}_{w_{ijk}=v}\mathbb{1}_{z^V_{ijk}=t})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ijk}+\mathbb{1}_{z^V_{ijk}=t})}\\
=&\frac{\Gamma(\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}+1)\prod_{t\neq t^V}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^t_{z^V_{ij,-k},z^H_{ij}})}{(T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)\Gamma(T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)}\\
&\times\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t^V}_{y^V=1,-ijk}+\mathbb{1}_{w_{ijk}=v})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^V}_{y^V=1,-ijk}+1)}\prod_{t\neq t^V}\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ijk})}\\
=&\frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}\frac{\prod_{t=1}^T\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^t_{z^V_{ij,-k},z^H_{ij}})}{\Gamma(T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)}\\
&\times\frac{\Gamma(\beta^V_v+n^{w^*,t^V}_{y^V=1,-ijk}+1)\prod_{v\neq w^*}\Gamma(\beta^V_v+n^{v,t^V}_{y^V=1,-ijk})}{(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^V}_{y^V=1,-ijk})\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t^V}_{y^V=1,-ijk})}\\
&\times\prod_{t\neq t^V}\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ijk})}\\
\propto&\frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}\\
&\times\frac{\beta^V_v+n^{w^*,t^V}_{y^V=1,-ijk}}{\sum_{v=1}^V\beta^V_v+n^{\cdot,t^V}_{y^V=1,-ijk}}\prod_{t=1}^T\frac{\prod_{v=1}^V\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^V\beta^V_v+n^{\cdot,t}_{y^V=1,-ijk})}\\
\propto&\frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^T\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}\times\frac{\beta^V_v+n^{w^*,t^V}_{y^V=1,-ijk}}{\sum_{v=1}^V\beta^V_v+n^{\cdot,t^V}_{y^V=1,-ijk}}
\end{aligned}
$$

La seconda è data da:

$$
p(z^V_{ijk}=t^V|y^V_{ijk}=1,x_{ij}=0,z^V_{ijk}=t^V,\ldots)
$$

$$
\begin{aligned}
&\propto \frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^{t}_{z^V_{ij,-k},z^H_{ij}}+\mathbb{1}_{z^V_{ijk}=t})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij})}\\
&\quad\times\prod_{t=1}^{T}\frac{\prod_{v=1}^{V}\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ijk}+\mathbb{1}_{w_{ijk}=v}\mathbb{1}_{z^*_{ij}=t})}{\Gamma(\sum_{v=1}^{V}\beta^V_v+n^{\cdot,t}_{y^V=1,-ijk}+\mathbb{1}_{z^*_{ij}=t})}\\
&\propto \frac{\Gamma(\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}+1)\prod_{t\neq t^V}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^{t}_{z^V_{ij,-k},z^H_{ij}})}{(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)}\\
&= \frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}\frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^{t}_{z^V_{ij,-k},z^H_{ij}})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)}\\
&\propto \frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}
\end{aligned}
$$

La terza coincide con la seconda ed è data da:

$$
\begin{aligned}
&p(z^V_{ijk}=t^V|y^V_{ijk}=0,z^V_{ijk}=t^V,\dots)\\
&\propto \frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^{t}_{z^V_{ij,-k},z^H_{ij}}+\mathbb{1}_{z^V_{ijk}=t})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij})}\\
&\quad\times\prod_{t=1}^{T}\frac{\prod_{v=1}^{V}\Gamma(\beta^V_v+n^{v,t}_{y^V=1,-ijk})}{\Gamma(\sum_{v=1}^{V}\beta^V_v+n^{\cdot,t}_{y^V=1,-ijk})}\\
&\propto \frac{\Gamma(\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}+1)\prod_{t\neq t^V}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^{t}_{z^V_{ij,-k},z^H_{ij}})}{(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)}\\
&= \frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}\frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ij,t}\alpha_t+n^{t}_{z^V_{ij,-k},z^H_{ij}})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1)}\\
&\propto \frac{\alpha_0+\lambda_{ij,t^V}\alpha_{t^V}+n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0+\sum_{t=1}^{T}\lambda_{ij,t}\alpha_t+N_{ij}+L_{ij}-1}
\end{aligned}
$$

Distribuzione del topic di un hashtag

La distribuzione di z^H_{ijk} date tutte le altre variabili è proporzionale al prodotto dei blocchi (6), (8):

$$
\begin{aligned}
&p(z^H_{ijk}|\dots)\\
&\propto\prod_{u=1}^{U}\prod_{d=1}^{D_u}\frac{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ud,t}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(\alpha_0+\lambda_{ud,t}\alpha_t+n^{t}_{z^V_{ud},z^H_{ud}})}{\Gamma(T\alpha_0+\sum_{t=1}^{T}\lambda_{ud,t}\alpha_t+N_{ud}+L_{ud})}
\end{aligned}
$$

$$\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n_{y^H=1}^{h,t})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n_{y^H=1}^{\cdot,t})}$$
$$\propto \frac{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ij,t}\alpha_t + n_{z_{ij}^V, z_{ij}^H}^{t})}{\Gamma(T\alpha_0 + \sum_{t=1}^{T} \lambda_{ij,t}\alpha_t + N_{ij} + L_{ij})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta_h^H + n_{y^H=1}^{h,t})}{\Gamma(\sum_{h=1}^{H} \beta_h^H + n_{y^H=1}^{\cdot,t})}$$

Si considerino le seguenti scomposizioni dei conteggi; in questo caso si vuole isolare z_{ijk}^H:

$$n_{z_{ij}^V, z_{ij}^H}^{t} = \sum_{n=1}^{N_{ij}} \mathbb{1}_{z_{ijn}^V = t} + \sum_{l=1}^{L_{ij}} \mathbb{1}_{z_{ijl}^H = t} = \sum_{n=1}^{N_{ij}} \mathbb{1}_{z_{ijn}^V = t} + \sum_{l \neq k} \mathbb{1}_{z_{ijl}^H = t} + \mathbb{1}_{z_{ijk}^H = t}$$
$$= n_{z_{ij}^V, z_{ij,-k}^H}^{t} + \mathbb{1}_{z_{ijk}^H = t} = \begin{cases} n_{z_{ij}^V, z_{ij,-k}^H}^{t} + 1 & \text{se } z_{ijk}^H = t \\ n_{z_{ij}^V, z_{ij,-k}^H}^{t} & \text{se } z_{ijk}^H \neq t \end{cases}$$

$$n_{y^H=1}^{h,t} = \sum_{u=1}^{U} \sum_{d=1}^{D_u} \sum_{l=1}^{L_{ud}} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H=1} (\mathbb{1}_{z_{udl}^H=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z_{ud}^*=t} \mathbb{1}_{x_{ud}=0})$$
$$= n_{y^H=1,-ijk}^{h,t} + \mathbb{1}_{h_{ijk}=h} \mathbb{1}_{y_{ijk}^H=1} (\mathbb{1}_{z_{ijk}^H=t} \mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z_{ij}^*=t} \mathbb{1}_{x_{ij}=0})$$
$$= \begin{cases} n_{y^H=1,-ijk}^{h,t} + \mathbb{1}_{h_{ijk}=h} \mathbb{1}_{z_{ijk}^H=t} & \text{se } y_{ijk}^H = 1,\ x_{ij} = 1 \\ n_{y^H=1,-ijk}^{h,t} + \mathbb{1}_{h_{ijk}=h} \mathbb{1}_{z_{ij}^*=t} & \text{se } y_{ijk}^H = 1,\ x_{ij} = 0 \\ n_{y^H=1,-ijk}^{h,t} & \text{se } y_{ijk}^H = 0 \end{cases}$$

$$n_{y^H=1}^{\cdot,t} = \sum_{h=1}^{H} n_{y^H=1}^{h,t} = n_{y^H=1,-ijk}^{\cdot,t} + \mathbb{1}_{y_{ijk}^H=1} (\mathbb{1}_{z_{ijk}^H=t} \mathbb{1}_{x_{ij}=1} + \mathbb{1}_{z_{ij}^*=t} \mathbb{1}_{x_{ij}=0})$$
$$= \begin{cases} n_{y^H=1,-ijk}^{\cdot,t} + \mathbb{1}_{z_{ijk}^H=t} & \text{se } y_{ijk}^H = 1,\ x_{ij} = 1 \\ n_{y^H=1,-ijk}^{\cdot,t} + \mathbb{1}_{z_{ij}^*=t} & \text{se } y_{ijk}^H = 1,\ x_{ij} = 0 \\ n_{y^H=1,-ijk}^{\cdot,t} & \text{se } y_{ijk}^H = 0 \end{cases}$$

I conteggi $n_{y^H=1}^{h,t}$ e $n_{y^H=1}^{\cdot,t}$ dipendono da z_{ijk}^H solo se $y_{ijk}^H = 1$ e $x_{ij} = 1$, è quindi possibile distinguere tre casi, $(y_{ijk}^H, x_{ij}) = (1,1)$, $(y_{ijk}^H, x_{ij}) = (1,0)$ e $y_{ijk}^H = 0$, quando si calcola la probabilità di $z_{ijk}^H = t^H$ date tutte le altre variabili; la prima è data da:

$$p(z_{ijk}^H = t^H | y_{ijk}^H = 1, x_{ij} = 1, z_{ijk}^H = t^H, h_{ijk} = h^*, \ldots)$$

$$\propto \frac{\alpha_0 + \lambda_{ij,t^H}\alpha_{t^H} + n^{t^H}_{z^V_{ij},z^H_{ij,-k}}}{T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1} \times \frac{\beta^H_h + n^{h^*,t^H}_{y^H=1,-ijk}}{\sum_{h=1}^{H}\beta^H_h + n^{\cdot,t^H}_{y^H=1,-ijk}}$$

La seconda è data da:

$$p(z^H_{ijk} = t^H | y^V_{ijk} = 1, x_{ij} = 0, z^H_{ijk} = t^H, \ldots)$$
$$\propto \frac{\alpha_0 + \lambda_{ij,t^H}\alpha_{t^H} + n^{t^H}_{z^V_{ij},z^H_{ij,-k}}}{T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1}$$

La terza coincide con la seconda ed è data da:

$$p(z^V_{ijk} = t^V | y^V_{ijk} = 0, z^H_{ijk} = t^H, \ldots)$$
$$\propto \frac{\alpha_0 + \lambda_{ij,t^H}\alpha_{t^H} + n^{t^H}_{z^V_{ij},z^H_{ij,-k}}}{T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1}$$

Il procedimento è stato omesso poiché coincide con quello esposto nella sezione precedente.

4.2.5 Stime dei Parametri

In questo caso, e più in generale per un qualsiasi *topic model*, si preferisce adottare il *Collapsed Gibbs Sampler* non perché i parametri del modello, **par**, non sono d'interesse, ma perché questo algoritmo risulta essere molto più efficiente rispetto al *Gibbs Sampler*. Sfruttando la scelta di utilizzare distribuzioni a priori coniugate, è possibile determinare la distribuzione di probabilità di ogni parametro in **par** e ottenere una sua stima utilizzando la formula del valore atteso della distribuzione identificata.

Il procedimento per ottenere le stime dei parametri è essenzialmente sempre lo stesso e consiste in:

1. considerare solo i blocchi della distribuzione congiunta p_{mod}[4] in cui compare il parametro di cui si vuole calcolare la stima;
2. determinare la distribuzione di probabilità del parametro date tutte le altre variabili del modello individuandone il nucleo;
3. ottenere una stima del parametro utilizzando la formula del valore atteso della distribuzione identificata nel punto precedente.

[4]Si ricordi che p_{mod} è la distribuzione congiunta delle variabili latenti **lat**, dei parametri con una distribuzione a priori associata, **par**, e delle variabili osservate **oss**.

La distribuzione di probabilità condizionata di ogni parametro segue la stessa distribuzione della sua distribuzioni a priori; più nello specifico, i parametri delle distribuzioni identificate dipendono sia dai parametri fissati delle distribuzioni a priori sia dalle altre variabili del modello.

Stima di δ

La distribuzione di δ date tutte le altre variabili è data da:

$$\begin{aligned}
p(\delta|\dots) \propto & p(\delta|b_1^\delta, b_2^\delta) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{t=1}^{T} p(\lambda_{ud,t}|\delta) \\
= & \frac{\Gamma(b_1^\delta + b_2^\delta)}{\Gamma(b_1^\delta)\Gamma(b_2^\delta)}\delta^{b_1^\delta-1}(1-\delta)^{b_2^\delta-1} \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{t=1}^{T} \delta^{\lambda_{ud,t}}(1-\delta)^{1-\lambda_{ud,t}} \\
\propto & \delta^{b_1^\delta-1}(1-\delta)^{b_2^\delta-1}\delta^{\sum_{udt}\mathbb{1}_{\lambda_{ud,t}=1}}(1-\delta)^{\sum_{udt}\mathbb{1}_{\lambda_{ud,t}=0}} \\
= & \delta^{b_1^\delta+\sum_{udt}\mathbb{1}_{\lambda_{ud,t}=1}-1}(1-\delta)^{b_2^\delta+\sum_{udt}\mathbb{1}_{\lambda_{ud,t}=0}-1} \\
= & \delta^{b_1^\delta+n_\lambda^1-1}(1-\delta)^{b_2^\delta+n_\lambda^0-1}
\end{aligned}$$

Si ottiene il nucleo di una distribuzione Beta con parametri $b_1^\delta + n_\lambda^1$ e $b_2^\delta + n_\lambda^0$, da cui

$$\delta|\dots \sim Beta(b_1^\delta + n_\lambda^1, b_2^\delta + n_\lambda^0)$$

Essendo $n_\lambda^1 + n_\lambda^0 = DT$, una stima di δ è data da

$$\delta = \frac{b_1^\delta + n_\lambda^1}{b_1^\delta + n_\lambda^1 + b_2^\delta + n_\lambda^0} = \frac{b_1^\delta + n_\lambda^1}{b_1^\delta + b_2^\delta + DT}$$

Stima di π^V

La distribuzione di π^V date tutte le altre variabili è data da:

$$\begin{aligned}
p(\pi^V|\dots) \propto & p(\pi^V|b_1^V, b_2^V) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{n=1}^{N_{ud}} p(y_{udn}^V|\pi^V) \\
= & \frac{\Gamma(b_1^V + b_2^V)}{\Gamma(b_1^V)\Gamma(b_2^V)}(\pi^V)^{b_1^V-1}(1-\pi^V)^{b_2^V-1} \prod_{udn}(\pi^V)^{y_{udn}^V}(1-\pi^V)^{1-y_{udn}^V} \\
\propto & (\pi^V)^{b_1^V-1}(1-\pi^V)^{b_2^V-1}(\pi^V)^{\sum_{udn}\mathbb{1}_{y_{udn}^V=1}}(1-\pi^V)^{\sum_{udn}\mathbb{1}_{y_{udn}^V=0}} \\
= & (\pi^V)^{b_1^V+\sum_{udn}\mathbb{1}_{y_{udn}^V=1}-1}(1-\pi^V)^{b_2^V+\sum_{udn}\mathbb{1}_{y_{udn}^V=0}-1} \\
= & (\pi^V)^{b_1^V+n_{y^V}^1-1}(1-\pi^V)^{b_2^V+n_{y^V}^0-1}
\end{aligned}$$

Si ottiene il nucleo di una distribuzione Beta con parametri $b_1^V + n_{y^V}^1$ e $b_2^V + n_{y^V}^0$, da cui

$$\pi^V | \ldots \sim Beta(b_1^V + n_{y^V}^1, b_2^V + n_{y^V}^0)$$

Essendo $n_{y^V}^1 + b_2^V + n_{y^V}^0 = N$, una stima di π^V è data da

$$\pi^V = \frac{b_1^V + n_{y^V}^1}{b_1^V + n_{y^V}^1 + b_2^V + n_{y^V}^0} = \frac{b_1^V + n_{y^V}^1}{b_1^V + b_2^V + N}$$

Stima di π^H

La distribuzione di π^H date tutte le altre variabili è data da:

$$\begin{aligned}
p(\pi^H | \ldots) \propto & p(\pi^H | b_1^H, b_2^H) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{l=1}^{L_{ud}} p(y_{udl}^H | \pi^H) \\
= & \frac{\Gamma(b_1^H + b_2^H)}{\Gamma(b_1^H)\Gamma(b_2^H)} (\pi^H)^{b_1^H - 1} (1 - \pi^H)^{b_2^H - 1} \prod_{udl} (\pi^H)^{y_{udl}^H} (1 - \pi^H)^{1 - y_{udl}^H} \\
\propto & (\pi^H)^{b_1^H - 1} (1 - \pi^H)^{b_2^H - 1} (\pi^H)^{\sum_{udl} \mathbb{1}_{y_{udl}^H = 1}} (1 - \pi^H)^{\sum_{udl} \mathbb{1}_{y_{udl}^H = 0}} \\
= & (\pi^H)^{b_1^H + \sum_{udl} \mathbb{1}_{y_{udl}^H = 1} - 1} (1 - \pi^H)^{b_2^H + \sum_{udl} \mathbb{1}_{y_{udl}^H = 0} - 1} \\
= & (\pi^H)^{b_1^H + n_{y^H}^1 - 1} (1 - \pi^H)^{b_2^H + n_{y^H}^0 - 1}
\end{aligned}$$

Si ottiene il nucleo di una distribuzione Beta con parametri $b_1^H + n_{y^H}^1$ e $b_2^H + n_{y^H}^0$, da cui

$$\pi^H | \ldots \sim Beta(b_1^H + n_{y^H}^1, b_2^H + n_{y^H}^0)$$

Essendo $n_{y^H}^1 + n_{y^H}^0 = L$, una stima di π^H è data da

$$\pi^H = \frac{b_1^H + n_{y^H}^1}{b_1^H + n_{y^H}^1 + b_2^H + n_{y^H}^0} = \frac{b_1^H + n_{y^H}^1}{b_1^H + b_2^H + L}$$

Stima di $\pi_{1:U}^T$

La distribuzione di π_u^T, $u = 1, \ldots, U$, date tutte le altre variabili è data da:

$$\begin{aligned}
p(\pi_u^T | \ldots) \propto & \prod_{u'=1}^{U} p(\pi_{u'}^T | b_1^T, b_2^T) \prod_{d=1}^{D_{u'}} p(x_{u'd} | \pi_{u'}^T) \\
\propto & p(\pi_u^T | b_1^T, b_2^T) \prod_{d=1}^{D_u} p(x_{ud} | \pi_u^T)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(b_1^T + b_2^T)}{\Gamma(b_1^T)\Gamma(b_2^T)} (\pi_u^T)^{b_1^T - 1} (1 - \pi_u^T)^{b_2^T - 1} \prod_{d=1}^{D_u} (\pi_u^T)^{x_{ud}} (1 - \pi_u^T)^{1 - x_{ud}} \\
&\propto (\pi_u^T)^{b_1^T - 1} (1 - \pi_u^T)^{b_2^T - 1} (\pi_u^T)^{\sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1}} (1 - \pi_u^T)^{\sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0}} \\
&= (\pi_u^T)^{b_1^T + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1} - 1} (1 - \pi_u^T)^{b_2^T + \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0} - 1} \\
&= (\pi_u^T)^{b_1^T + n_{x_u}^1 - 1} (1 - \pi_u^T)^{b_2^T + n_{x_u}^0 - 1}
\end{aligned}
$$

Si ottiene il nucleo di una distribuzione Beta con parametri $b_1^T + n_{x_u}^1$ e $b_2^H + n_{x_u}^0$, da cui

$$\pi^H | \ldots \sim Beta(b_1^H + n_{x_u}^1, b_2^H + n_{x_u}^0)$$

Essendo $n_{x_u}^1 + n_{x_u}^0 = D_u$, una stima di π_u^T, $u = 1, \ldots, U$, è data da

$$\pi_u^T = \frac{b_1^T + n_{x_u}^1}{b_1^T + n_{x_u}^1 + b_2^T + n_{x_u}^0} = \frac{b_1^T + n_{x_u}^1}{b_1^T + b_2^T + D_u}$$

Stima di $\theta_{1:U}^*$

La distribuzione di $\boldsymbol{\theta}_u^*$, $u = 1, \ldots, U$, date tutte le altre variabili è data da:

$$
\begin{aligned}
p(\boldsymbol{\theta}_u^* | \ldots) &\propto \prod_{u'=1}^{U} p(\boldsymbol{\theta}_{u'}^* | \boldsymbol{\alpha}^*) \prod_{d=1}^{D_{u'}} p(z_{u'd}^* | \boldsymbol{\theta}_{u'}^*) \\
&= p(\boldsymbol{\theta}_u^* | \boldsymbol{\alpha}^*) \prod_{d=1}^{D_u} p(z_{ud}^* | \boldsymbol{\theta}_u^*) \\
&= \frac{\Gamma(\sum_{t=1}^{T} \alpha_t^*)}{\prod_{t=1}^{T} \Gamma(\alpha_t^*)} \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} \prod_{d=1}^{D_u} (\theta_{u,t}^*)^{\mathbb{1}_{z_{ud}^* = t}} \\
&\propto \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* - 1} (\theta_{u,t}^*)^{\sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^* = t}} \\
&= \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* + \sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^* = t} - 1} \\
&= \prod_{t=1}^{T} (\theta_{u,t}^*)^{\alpha_t^* + n_{z_u^*}^t - 1}
\end{aligned}
$$

Si ottiene il nucleo di una distribuzione di Dirichlet di ordine T con parametro $T \times 1$ il cui t-mo elemento è dato da $\alpha_t^* + n_{z_u^*}^t$, da cui

$$\boldsymbol{\theta}_u^* | \ldots \sim Dir_T(\alpha^* + n_{z_u^*}^{1:T})$$

dove $n^{1:T}_{z^*_u} = (n^1_{z^*_u} \cdots n^T_{z^*_u})$. Essendo $\sum_{t=1}^T n^t_{z^*_u} = D_u$, una stima del t-mo elemento di $\boldsymbol{\theta}^*_u$ è data da

$$\theta^*_{u,t} = \frac{\alpha^*_t + n^t_{z^*_u}}{\sum_{t'=1}^T \left(\alpha^*_{t'} + n^{t'}_{z^*_u}\right)} = \frac{\alpha^*_t + n^t_{z^*_u}}{\sum_{t'=1}^T \alpha^*_{t'} + D_u}$$

per $u = 1, \ldots, U$, $t = 1, \ldots, T$.

Stima di $\theta_{1:D}$

La distribuzione di $\boldsymbol{\theta}_{ud}$, $u = 1, \ldots, U$, $d = 1, \ldots, D_u$, date tutte le altre variabili è data da:

$$\begin{aligned}
p(\boldsymbol{\theta}_{ud}|\ldots) \propto & \prod_{u'=1}^{U} \prod_{d'=1}^{D_{u'}} p(\boldsymbol{\theta}_{u'd'}|\boldsymbol{\lambda}_{u'd'}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{u'd'}} p(z^V_{u'd'n}|\boldsymbol{\theta}_{u'd'}) \prod_{l=1}^{L_{u'd'}} p(z^H_{u'd'l}|\boldsymbol{\theta}_{u'd'}) \\
= & p(\boldsymbol{\theta}_{ud}|\boldsymbol{\lambda}_{ud}, \boldsymbol{\alpha}, \alpha_0) \prod_{n=1}^{N_{ud}} p(z^V_{udn}|\boldsymbol{\theta}_{ud}) \prod_{l=1}^{L_{ud}} p(z^H_{udl}|\boldsymbol{\theta}_{ud}) \\
= & \frac{\Gamma(\sum_{t=1}^T (\alpha_0 + \lambda_{ud,t}\alpha_t))}{\prod_{t=1}^T \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \prod_{t=1}^{T} \theta^{\alpha_0 + \lambda_{ud,t}\alpha_t - 1}_{ud,t} \prod_{n=1}^{N_{ud}} \theta^{\mathbb{1}_{z^V_{udn}=t}}_{ud,t} \prod_{l=1}^{L_{ud}} \theta^{\mathbb{1}_{z^H_{udl}=t}}_{ud,t} \\
\propto & \prod_{t=1}^{T} \theta^{\alpha_0 + \lambda_{ud,t}\alpha_t + \sum_{n=1}^{N_{ud}} \mathbb{1}_{z^V_{udn}=t} + \sum_{l=1}^{L_{ud}} \mathbb{1}_{z^H_{udl}=t} - 1}_{ud,t} \\
= & \prod_{t=1}^{T} \theta^{\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud}, z^H_{ud}} - 1}_{ud,t}
\end{aligned}$$

Si ottiene il nucleo di una distribuzione di Dirichlet di ordine T con parametro $T \times 1$ il cui t-mo elemento è dato da $\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud}, z^H_{ud}}$, da cui

$$\boldsymbol{\theta}_{ud}|\ldots \sim Dir_T(\alpha_0 \mathbf{1}_T + \boldsymbol{\lambda}_{ud} \circ \boldsymbol{\alpha} + n^{1:T}_{z^V_{ud}, z^H_{ud}})$$

dove $\boldsymbol{\lambda}_{ud} = (\lambda_{ud,1} \cdots \lambda_{ud,T})$ e $n^{1:T}_{z^V_{ud}, z^H_{ud}} = (n^1_{z^V_{ud}, z^H_{ud}} \cdots n^T_{z^V_{ud}, z^H_{ud}})$.
Essendo $\sum_{t=1}^T n^t_{z^V_{ud}, z^H_{ud}} = N_{ud} + L_{ud}$, una stima del t-mo elemento di $\boldsymbol{\theta}_{ud}$ è data da

$$\theta_{ud,t} = \frac{\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud}, z^H_{ud}}}{\sum_{t'=1}^T \left(\alpha_0 + \lambda_{ud,t'}\alpha_{t'} + n^{t'}_{z^V_{ud}, z^H_{ud}}\right)} = \frac{\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud}, z^H_{ud}}}{T\alpha_0 + \sum_{t'=1}^T \lambda_{ud,t'}\alpha_{t'} + N_{ud} + L_{ud}}$$

per $u = 1, \ldots, U$, $d = 1, \ldots, D_u$, $t = 1, \ldots, T$.

Stima di $\phi_{1:T}$

La distribuzione di $\boldsymbol{\phi}_t$, $t = 1, \ldots, T$, date tutte le altre variabili è data da:

$$
\begin{aligned}
p(\boldsymbol{\phi}_t|\ldots) &\propto \prod_{t'=1}^{T} p(\boldsymbol{\phi}_{t'}|\boldsymbol{\beta}^V) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
&= \left(\prod_{t'=1}^{T} \frac{\Gamma(\sum_{v=1}^{V} \beta_v^V)}{\prod_{v=1}^{V} \Gamma(\beta_v^V)} \prod_{v=1}^{V} \phi_{t',v}^{\beta_v^V - 1} \right) \\
&\quad \times \left(\prod_{udn} \prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{x_{ud}=0}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=0}} \right) \\
&\propto \left(\prod_{t'=1}^{T} \prod_{v=1}^{V} \phi_{t',v}^{\beta_v^V - 1} \right) \left(\prod_{t'=1}^{T} \prod_{udn} \prod_{v=1}^{V} \phi_{t',v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{z^V_{udn}=t'} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \phi_{t',v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{z^*_{ud}=t'} \mathbb{1}_{x_{ud}=0}} \right) \\
&\propto \left(\prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \right) \left(\prod_{udn} \prod_{v=1}^{V} \phi_{t,v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{z^V_{udn}=t} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \phi_{t,v}^{\mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0}} \right) \\
&= \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V - 1} \phi_{t,v}^{\sum_{udn} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} (\mathbb{1}_{z^V_{udn}=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0})} \\
&= \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V + \sum_{udn} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} (\mathbb{1}_{z^V_{udn}=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0}) - 1} \\
&= \prod_{v=1}^{V} \phi_{t,v}^{\beta_v^V + n^{v,t}_{y^V=1} - 1}
\end{aligned}
$$

Si ottiene il nucleo di una distribuzione di Dirichlet di ordine V con parametro $V \times 1$ il cui v-mo elemento è dato da $\beta_v^V + n^{v,t}_{y^V=1}$, da cui

$$\boldsymbol{\phi}_t|\ldots \sim Dir_V(\beta^V + n^{v,1:T}_{y^V=1})$$

dove $n^{v,1:T}_{y^V=1} = (n^{v,1}_{y^V=1} \cdots n^{v,T}_{y^V=1})$. Sia $\sum_{v=1}^{V} n^{v,t}_{y^V=1} = n^{\cdot,t}_{y^V=1}$, una stima del v-mo elemento di $\boldsymbol{\phi}_t$ è data da

$$\phi_{t,v} = \frac{\beta_v^V + n^{v,t}_{y^V=1}}{\sum_{v'=1}^{V} \left(\beta_{v'}^V + n^{v',t}_{y^V=1} \right)} = \frac{\beta_v^V + n^{v,t}_{y^V=1}}{\sum_{v'=1}^{V} \beta_{v'}^V + n^{\cdot,t}_{y^V=1}}$$

per $t = 1, \ldots, T$, $v = 1, \ldots, V$.

Stima di $\psi_{1:T}$

La distribuzione di $\boldsymbol{\psi}_t$, $t = 1, \dots, T$, date tutte le altre variabili è data da:

$$
\begin{aligned}
p(\boldsymbol{\psi}_t|\dots) &\propto \prod_{t'=1}^{T} p(\boldsymbol{\psi}_{t'}|\boldsymbol{\beta}^H) \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{l=1}^{L_{ud}} p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}}) \\
&= \left(\prod_{t'=1}^{T} \frac{\Gamma(\sum_{h=1}^{H} \beta^H_h)}{\prod_{h=1}^{H} \Gamma(\beta^H_h)} \prod_{h=1}^{H} \psi_{t',h}^{\beta^H_h - 1} \right) \\
&\quad \times \left(\prod_{udl} \prod_{h=1}^{H} \psi_{z^H_{udl},h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \psi_{z^*_{ud},h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{x_{ud}=0}} (\psi^{\mathcal{B}}_h)^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=0}} \right) \\
&\propto \left(\prod_{t'=1}^{T} \prod_{h=1}^{H} \psi_{t',h}^{\beta^H_h - 1} \right) \left(\prod_{t'=1}^{T} \prod_{udl} \prod_{h=1}^{H} \psi_{t',h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{z^H_{udl}=t'} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \psi_{t',h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{z^*_{ud}=t'} \mathbb{1}_{x_{ud}=0}} \right) \\
&\propto \left(\prod_{h=1}^{H} \psi_{t,h}^{\beta^H_h - 1} \right) \left(\prod_{udl} \prod_{h=1}^{H} \psi_{t,h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{z^H_{udl}=t} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \psi_{t,h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0}} \right) \\
&= \prod_{h=1}^{H} \psi_{t,h}^{\beta^H_h - 1} \psi_{t,h}^{\sum_{udl} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} (\mathbb{1}_{z^H_{udl}=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0})} \\
&= \prod_{h=1}^{H} \psi_{t,h}^{\beta^H_h + \sum_{udl} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} (\mathbb{1}_{z^H_{udl}=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z^*_{ud}=t} \mathbb{1}_{x_{ud}=0}) - 1} \\
&= \prod_{h=1}^{H} \psi_{t,h}^{\beta^H_h + n^{h,t}_{y^H=1} - 1}
\end{aligned}
$$

Si ottiene il nucleo di una distribuzione di Dirichlet di ordine H con parametro $H \times 1$ il cui h-mo elemento è dato da $\beta^H_h + n^{h,t}_{y^H=1}$, da cui

$$\boldsymbol{\psi}_t|\dots \sim Dir_H(\beta^H + n^{h,1:T}_{y^H=1})$$

dove $n^{h,1:T}_{y^H=1} = (n^{h,1}_{y^H=1} \cdots n^{h,T}_{y^H=1})$. Sia $\sum_{h=1}^{H} n^{h,t}_{y^H=1} = n^{\cdot,t}_{y^H=1}$, una stima dell'$h$-mo elemento di $\boldsymbol{\psi}_t$ è data da

$$\psi_{t,h} = \frac{\beta^H_h + n^{h,t}_{y^H=1}}{\sum_{h'=1}^{H} \left(\beta^H_{h'} + n^{h',t}_{y^H=1} \right)} = \frac{\beta^H_h + n^{h,t}_{y^H=1}}{\sum_{h'=1}^{H} \beta^H_{h'} + n^{\cdot,t}_{y^H=1}}$$

per $t = 1, \dots, T$, $h = 1, \dots, H$.

Stima di $\phi^{\mathcal{B}}$

La distribuzione di $\boldsymbol{\phi}^{\mathcal{B}}$, $t = 1, \ldots, T$, date tutte le altre variabili è data da:

$$
\begin{aligned}
p(\boldsymbol{\phi}^{\mathcal{B}}|\ldots) \propto & p(\boldsymbol{\phi}^{\mathcal{B}}|\boldsymbol{\beta}^V) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{n=1}^{N_{ud}} p(w_{udn}|y^V_{udn}, z^V_{udn}, x_{ud}, z^*_{ud}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}) \\
= & \left(\frac{\Gamma(\sum_{v=1}^V \beta_v^V)}{\prod_{v=1}^V \Gamma(\beta_v^V)} \prod_{v=1}^V (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} \right) \\
& \times \left(\prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{n=1}^{N_{ud}}\prod_{v=1}^{V} \phi_{z^V_{udn},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=1}} \right. \\
& \left. \times \phi_{z^*_{ud},v}^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=1}\mathbb{1}_{x_{ud}=0}} (\phi_v^{\mathcal{B}})^{\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \right) \\
\propto & \prod_{v=1}^V (\phi_v^{\mathcal{B}})^{\beta_v^V - 1} (\phi_v^{\mathcal{B}})^{\sum_{udn}\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0}} \\
= & \prod_{v=1}^V (\phi_v^{\mathcal{B}})^{\beta_v^V + \sum_{udn}\mathbb{1}_{w_{udn}=v}\mathbb{1}_{y^V_{udn}=0} - 1} \\
= & \prod_{v=1}^V (\phi_v^{\mathcal{B}})^{\beta_v^V + n^v_{y^V=0} - 1}
\end{aligned}
$$

Si ottiene il nucleo di una distribuzione di Dirichlet di ordine V con parametro $V \times 1$ il cui v-mo elemento è dato da $\beta_v^V + n^v_{y^V=0}$, da cui

$$\boldsymbol{\phi}^{\mathcal{B}}|\ldots \sim Dir_V(\beta^V + n^{1:V}_{y^V=0})$$

dove $n^{1:V}_{y^V=0} = (n^1_{y^V=0} \cdots n^V_{y^V=0})$. Sia $\sum_{v=1}^V n^v_{y^V=0} = n^{\cdot}_{y^V=0}$, una stima del v-mo elemento di $\boldsymbol{\phi}^{\mathcal{B}}$ è data da

$$\phi_v^{\mathcal{B}} = \frac{\beta_v^V + n^v_{y^V=0}}{\sum_{v'=1}^V \left(\beta_{v'}^V + n^{v'}_{y^V=0}\right)} = \frac{\beta_v^V + n^v_{y^V=0}}{\sum_{v'=1}^V \beta_{v'}^V + n^{\cdot}_{y^V=0}}$$

per $v = 1, \ldots, V$.

Stima di $\psi^{\mathcal{B}}$

La distribuzione di $\boldsymbol{\psi}^{\mathcal{B}}$, $t = 1, \ldots, T$, date tutte le altre variabili è data da:

$$p(\boldsymbol{\psi}^{\mathcal{B}}|\ldots) \propto p(\boldsymbol{\psi}^{\mathcal{B}}|\boldsymbol{\beta}^H) \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{l=1}^{L_{ud}} p(h_{udl}|y^H_{udl}, z^H_{udl}, x_{ud}, z^*_{ud}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\psi}^{\mathcal{B}})$$

$$
\begin{aligned}
&= \left(\frac{\Gamma(\sum_{h=1}^{H} \beta_h^H)}{\prod_{h=1}^{H} \Gamma(\beta_h^H)} \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} \right) \\
&\quad \times \left(\prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} \prod_{h=1}^{H} \psi_{z_{udl}^H, h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H=1} \mathbb{1}_{x_{ud}=1}} \right. \\
&\quad \left. \times \psi_{z_{ud}^*, h}^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H=1} \mathbb{1}_{x_{ud}=0}} (\psi_h^{\mathcal{B}})^{\mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H=0}} \right) \\
&\propto \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H - 1} (\psi_h^{\mathcal{B}})^{\sum_{udl} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H=0}} \\
&= \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H + \sum_{udl} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y_{udl}^H=0} - 1} \\
&= \prod_{h=1}^{H} (\psi_h^{\mathcal{B}})^{\beta_h^H + n_{y^H=0}^h - 1}
\end{aligned}
$$

Si ottiene il nucleo di una distribuzione di Dirichlet di ordine H con parametro $H \times 1$ il cui h-mo elemento è dato da $\beta_h^H + n_{y^H=0}^h$, da cui

$$\boldsymbol{\psi}^{\mathcal{B}} | \ldots \sim Dir_H(\beta^H + n_{y^H=0}^{1:H})$$

dove $n_{y^H=0}^{1:H} = {}^{(n_{y^H=0}^1 \cdots n_{y^H=0}^H)}$. Sia $\sum_{h=1}^{H} n_{y^H=0}^h = n_{y^H=0}^{\cdot}$, una stima dell'$h$-mo elemento di $\boldsymbol{\psi}^{\mathcal{B}}$ è data da

$$\psi_h^{\mathcal{B}} = \frac{\beta_h^H + n_{y^H=0}^h}{\sum_{h'=1}^{H} \left(\beta_{h'}^H + n_{y^H=0}^{h'} \right)} = \frac{\beta_h^H + n_{y^H=0}^h}{\sum_{h'=1}^{H} \beta_{h'}^H + n_{y^H=0}^{\cdot}}$$

per $h = 1, \ldots, H$.

4.3 Blocked Collapsed Gibbs Sampler

Il *Collapsed Gibbs Sampler* costruito nella sezione 4.2 aggiorna un'unica variabile aleatoria alla volta estraendo nuovi valori delle *full conditional probabilities*: in questo caso l'algoritmo si dice *pointwise*. Alternativamente, è possibile aggiornare più variabili aleatorie ottenendo quindi un *Blocked Collapsed Gibbs Sampler* (Gao & Johnson, 2008).

A ogni iterazione del *Pointwise Collapsed Gibbs Sampler* introdotto in precedenza è necessario

- aggiornare il tipo, x_{ud}, e il topic principale, z^*_{ud}, di ogni documento ud;
- aggiornare l'origine, y^V_{udn}, e il topic, z^V_{udn}, di ogni parola udn;
- aggiornare l'origine, y^H_{udl}, e il topic, z^H_{udl}, di ogni parola udl.

Si propone quindi un *Blocked Collapsed Gibbs Sampler* in cui le variabili relative a uno stesso documento, a una stessa parola e a uno stesso *hashtag* sono aggiornate congiuntamente; il procedimento per costruire il nuovo algoritmo coincide essenzialmente con quello utilizzato per il *Pointwise Collapsed Gibbs Sampler*, tuttavia al posto di derivare le *full conditional probabilities* è necessario determinare le distribuzioni congiunte condizionate di (x_{ud}, z^*_{ud}), (y^V_{udn}, z^V_{udn}) e (y^H_{udl}, z^H_{udl}) date tutte le altre variabili aleatorie.

Distribuzione congiunta del tipo e del topic principale di un documento

La distribuzione di (x_{ij}, z^*_{ij}) date tutte le altre variabili è proporzionale al prodotto dei blocchi (4), (5), (8) e (10):

$$
\begin{aligned}
&p(x_{ij}, z^*_{ij} | \dots) \\
&\propto \prod_{u=1}^{U} \frac{\Gamma(b_1^T + n^1_{x_u})\Gamma(b_2^T + n^0_{x_u})}{\Gamma(b_1^T + b_2^T + D_u)} \times \prod_{u=1}^{U} \frac{\prod_{t=1}^{T} \Gamma(\alpha^*_t + n^t_{z^*_u})}{\Gamma(\sum_{t=1}^{T} \alpha^*_t + D_u)} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta^V_v + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V} \beta^V_v + n^{\cdot,t}_{y^V=1})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta^H_h + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^{H} \beta^H_h + n^{\cdot,t}_{y^H=1})} \\
&\propto \frac{\Gamma(b_1^T + n^1_{x_i})\Gamma(b_2^T + n^0_{x_i})}{\Gamma(b_1^T + b_2^T + D_i)} \times \frac{\prod_{t=1}^{T} \Gamma(\alpha^*_t + n^t_{z^*_i})}{\Gamma(\sum_{t=1}^{T} \alpha^*_t + D_i)} \\
&\quad \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta^V_v + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V} \beta^V_v + n^{\cdot,t}_{y^V=1})} \times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \Gamma(\beta^H_h + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^{H} \beta^H_h + n^{\cdot,t}_{y^H=1})}
\end{aligned}
$$

Sfruttando le quantità definite nelle sezioni precedenti e seguendo un procedimento molto simile, si possono facilmente determinare le probabilità di $(x_{ij} = 1, z^*_{ij} = t^*)$ e $(x_{ij} = 0, z^*_{ij} = t^*)$; la prima è data da:

$$
\begin{aligned}
&p(x_{ij} = 1, z^*_{ij} = t^* | \dots) \\
&\propto \frac{b_1^T + n^1_{x_i,-j}}{b_1^T + b_2^T + D_i - 1} \times \frac{\alpha^*_{t^*} + n^{t^*}_{z^*_i,-j}}{\sum_{t=1}^{T} \alpha^*_t + D_i - 1}
\end{aligned}
$$

$$\times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} (\beta_v^V + n^{v,t}_{y^V=1,-ij} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} \mathbb{1}_{z^V_{ijn}=t} - 1} (\beta_v^V + n^{\cdot,t}_{y^V=1,-ij} + q)}$$

$$\times \prod_{t=1}^{T} \frac{\prod_{h=1}^{H} \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} (\beta_h^H + n^{h,t}_{y^H=1,-ij} + q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} \mathbb{1}_{z^H_{ijl}=t} - 1} (\beta_h^H + n^{\cdot,t}_{y^H=1,-ij} + q)}$$

La seconda è data da:

$$\begin{aligned}
&p(x_{ij} = 0, z^*_{ij} = t^* | \ldots) \\
\propto &\frac{b_2^T + n^0_{x_i,-j}}{b_1^T + b_2^T + D_i - 1} \times \frac{\alpha^*_{t^*} + n^{t^*}_{z^*_i,-j}}{\sum_{t=1}^{T} \alpha^*_t + D_i - 1} \\
&\times \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1} - 1} (\beta_v^V + n^{v,t^*}_{y^V=1,-ij} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} - 1} (\sum_{v=1}^{V} \beta_v^V + n^{\cdot,t^*}_{y^V=1,-ij} + q)} \\
&\times \frac{\prod_{h=1}^{H} \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y^H_{ijl}=1} - 1} (\beta_h^H + n^{h,t^*}_{y^H=1,-ij} + q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y^H_{ijl}=1} - 1} (\sum_{h=1}^{H} \beta_h^H + n^{\cdot,t^*}_{y^H=1,-ij} + q)}
\end{aligned}$$

In questo caso esistono $2T$ possibili coppie di valori osservabili: T coppie con $x_{ij} = 1$ e altre T con $x_{ij} = 0$.

Distribuzione congiunta dell'origine e del topic di una parola

La distribuzione di (y^V_{ijk}, z^V_{ijk}) date tutte le altre variabili è proporzionale al prodotto dei blocchi (2), (6), (7) e (8):

$$\begin{aligned}
&p(y^V_{ijk}, z^V_{ijk} | \ldots) \\
\propto &\frac{\Gamma(b_1^V + n^1_{y^V}) \Gamma(b_2^V + n^0_{y^V})}{\Gamma(b_1^V + b_2^V + N)} \\
&\times \prod_{u=1}^{U} \prod_{d=1}^{D_u} \frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T} \lambda_{ud,t} \alpha_t)}{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t} \alpha_t)} \frac{\prod_{t=1}^{T} \Gamma(\alpha_0 + \lambda_{ud,t} \alpha_t + n^t_{z^V_{ud}, z^H_{ud}})}{\Gamma(T\alpha_0 + \sum_{t=1}^{T} \lambda_{ud,t} \alpha_t + N_{ud} + L_{ud})} \\
&\times \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^v_{y^V=0})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot}_{y^V=0})} \times \prod_{t=1}^{T} \frac{\prod_{v=1}^{V} \Gamma(\beta_v^V + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V} \beta_v^V + n^{\cdot,t}_{y^V=1})}
\end{aligned}$$

$$\propto \frac{\Gamma(b_1^V + n^1_{y^V})\Gamma(b_2^V + n^0_{y^V})}{\Gamma(b_1^V + b_2^V + N)} \times \frac{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ij,t}\alpha_t + n^t_{z^V_{ij}, z^H_{ij}})}{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij})}$$
$$\times \frac{\prod_{v=1}^{V}\Gamma(\beta_v^V + n^v_{y^V=0})}{\Gamma(\sum_{v=1}^{V}\beta_v^V + n^{\cdot}_{y^V=0})} \times \prod_{t=1}^{T}\frac{\prod_{v=1}^{V}\Gamma(\beta_v^V + n^{v,t}_{y^V=1})}{\Gamma(\sum_{v=1}^{V}\beta_v^V + n^{\cdot,t}_{y^V=1})}$$

Sfruttando le quantità definite nelle sezioni precedenti e seguendo un procedimento molto simile, si possono facilmente determinare le probabilità di $(y^V_{ijk} = 0, z^V_{ijk} = t^V)$ e $(z^V_{ijk} = 1, z^V_{ijk} = t^V)$; la prima è data da:

$$p(y^V_{ijk} = 0, z^V_{ijk} = t^V | w_{ijk} = w^*, \ldots)$$
$$\propto \frac{b_2^V + n^0_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \times \frac{\alpha_0 + \lambda_{ij,t^V}\alpha_{t^V} + n^{t^V}_{z^V_{ij,-k}, z^H_{ij}}}{T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1}$$
$$\times \frac{\beta^V_{w^*} + n^{w^*}_{y^V=0,-ijk}}{\sum_{v=1}^{V}\beta_v^V + n^{\cdot}_{y^V=0,-ijk}}$$

È possibile distinguere due casi, $x_{ij} = 1$ e $x_{ij} = 0$, quando si calcola la probabilità di $(y^V_{ijk} = 1, z^V_{ijk} = t^V)$; nel primo caso è data da:

$$p(y^V_{ijk} = 1, z^V_{ijk} = t^V | x_{ij} = 1, w_{ijk} = w^*, \ldots)$$
$$\propto \frac{b_1^V + n^1_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \times \frac{\alpha_0 + \lambda_{ij,t^V}\alpha_{t^V} + n^{t^V}_{z^V_{ij,-k}, z^H_{ij}}}{T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1}$$
$$\times \frac{\beta^V_{w^*} + n^{w^*,t^V}_{y^V=1,-ijk}}{\sum_{v=1}^{V}\beta_v^V + n^{\cdot,t^V}_{y^V=1,-ijk}}$$

Nel secondo caso è data da:

$$p(y^V_{ijk} = 1, z^V_{ijk} = t^V | x_{ij} = 0, w_{ijk} = w^*, \ldots)$$
$$\propto \frac{b_1^V + n^1_{y^V,-ijk}}{b_1^V + b_2^V + N - 1} \times \frac{\alpha_0 + \lambda_{ij,t^V}\alpha_{t^V} + n^{t^V}_{z^V_{ij,-k}, z^H_{ij}}}{T\alpha_0 + \sum_{t=1}^{T}\lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1}$$
$$\times \frac{\beta^V_{w^*} + n^{w^*,t^*}_{y^V=1,-ijk}}{\sum_{v=1}^{V}\beta_v^V + n^{\cdot,t^*}_{y^V=1,-ijk}}$$

In questo caso esistono $2T$ possibili coppie di valori osservabili: T coppie con $y^V_{ijk} = 1$ e altre T con $y^V_{ijk} = 0$.

Distribuzione congiunta dell'origine e del topic di un hashtag

La distribuzione di (y^H_{ijk}, z^H_{ijk}) date tutte le altre variabili è proporzionale al prodotto dei blocchi (2), (6), (9) e (10):

$$
\begin{aligned}
&p(y^H_{ijk}, z^H_{ijk}| \dots) \\
\propto &\frac{\Gamma(b_1^H + n^1_{y^H})\Gamma(b_2^H + n^0_{y^H})}{\Gamma(b_1^H + b_2^H + L)} \\
&\times \prod_{u=1}^{U}\prod_{d=1}^{D_u} \frac{\Gamma(T\alpha_0 + \sum_{t=1}^T \lambda_{ud,t}\alpha_t)}{\prod_{t=1}^T \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)} \frac{\prod_{t=1}^T \Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud}, z^H_{ud}})}{\Gamma(T\alpha_0 + \sum_{t=1}^T \lambda_{ud,t}\alpha_t + N_{ud} + L_{ud})} \\
&\times \frac{\prod_{h=1}^H \Gamma(\beta_h^H + n^h_{y^H=0})}{\Gamma(\sum_{h=1}^H \beta_h^H + n^{\cdot}_{y^H=0})} \times \prod_{t=1}^T \frac{\prod_{h=1}^H \Gamma(\beta_h^H + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^H \beta_h^H + n^{\cdot,t}_{y^H=1})} \\
\propto &\frac{\Gamma(b_1^H + n^1_{y^H})\Gamma(b_2^H + n^0_{y^H})}{\Gamma(b_1^H + b_2^H + L)} \times \frac{\prod_{t=1}^T \Gamma(\alpha_0 + \lambda_{ij,t}\alpha_t + n^t_{z^V_{ij}, z^H_{ij}})}{\Gamma(T\alpha_0 + \sum_{t=1}^T \lambda_{ij,t}\alpha_t + N_{ij} + L_{ij})} \\
&\times \frac{\prod_{h=1}^H \Gamma(\beta_h^H + n^h_{y^H=0})}{\Gamma(\sum_{h=1}^H \beta_h^H + n^{\cdot}_{y^H=0})} \times \prod_{t=1}^T \frac{\prod_{v=1}^V \Gamma(\beta_h^H + n^{h,t}_{y^H=1})}{\Gamma(\sum_{h=1}^H \beta_h^H + n^{\cdot,t}_{y^H=1})}
\end{aligned}
$$

Sfruttando le quantità definite nelle sezioni precedenti e seguendo un procedimento molto simile, si possono facilmente determinare le probabilità di $(y^H_{ijk} = 0, z^H_{ijk} = t^H)$ e $(y^H_{ijk} = 1, z^H_{ijk} = t^H)$; la prima è data da:

$$
\begin{aligned}
&p(y^H_{ijk} = 0, z^H_{ijk} = t^H | h_{ijk} = h^*, \dots) \\
\propto &\frac{b_2^H + n^0_{y^H,-ijk}}{b_1^H + b_2^H + L - 1} \times \frac{\alpha_0 + \lambda_{ij,t^H}\alpha_{t^H} + n^{t^H}_{z^V_{ij}, z^H_{ij,-k}}}{T\alpha_0 + \sum_{t=1}^T \lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1} \\
&\times \frac{\beta^H_{h^*} + n^{h^*}_{y^H=0,-ijk}}{\sum_{h=1}^H \beta_h^H + n^{\cdot}_{y^H=0,-ijk}}
\end{aligned}
$$

È possibile distinguere due casi, $x_{ij} = 1$ e $x_{ij} = 0$, quando si calcola la probabilità di $(y^H_{ijk} = 1, z^H_{ijk} = t^H)$; nel primo caso è data da:

$$
\begin{aligned}
&p(y^H_{ijk} = 1, z^H_{ijk} = t^H | x_{ij} = 1, h_{ijk} = h^*, \dots) \\
\propto &\frac{b_1^H + n^1_{y^H,-ijk}}{b_1^H + b_2^H + L - 1} \times \frac{\alpha_0 + \lambda_{ij,t^H}\alpha_{t^H} + n^{t^H}_{z^V_{ij}, z^H_{ij,-k}}}{T\alpha_0 + \sum_{t=1}^T \lambda_{ij,y}\alpha_t + N_{ij} + L_{ij} - 1} \\
&\times \frac{\beta^V_{h^*} + n^{h^*,t^H}_{y^H=1,-ijk}}{\sum_{h=1}^H \beta_v^H + n^{\cdot,t^H}_{y^H=1,-ijk}}
\end{aligned}
$$

Nel secondo caso è data da:

$$\begin{aligned} & p(y^H_{ijk} = 1, z^H_{ijk} = t^H | x_{ij} = 0, h_{ijk} = h^*, x_{ij} = t^* \ldots) \\ \propto & \frac{b^H_1 + n^1_{y^H,-ijk}}{b^H_1 + b^H_2 + L - 1} \times \frac{\alpha_0 + \lambda_{ij,t^H}\alpha_{t^H} + n^{t^H}_{z^V_{ij},z^H_{ij,-k}}}{T\alpha_0 + \sum_{t=1}^{T} \lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1} \\ & \times \frac{\beta^H_{h^*} + n^{h^*,t^*}_{y^H=1,-ijk}}{\sum_{h=1}^{H} \beta^H_v + n^{\cdot,t^*}_{y^H=1,-ijk}} \end{aligned}$$

In questo caso esistono $2T$ possibili coppie di valori osservabili: T coppie con $y^H_{ijk} = 1$ e altre T con $y^H_{ijk} = 0$.

4.4 Casi Particolari dell'Algoritmo

Fissando opportunamente le variabili latenti in **par**, il *Collapsed Gibbs Sampler* del modello proposto in questa tesi può essere utilizzato per effettuare l'inferenza della *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA*. In particolare, riprendendo la terminologia introdotta nella sezione 3.6, basta tenere fissate le variabili latenti presenti nel modello esteso ma non in quello ristretto ed aggiornare esclusivamente le variabili latenti in **lat** presenti nel modello ristretto; analogamente, i parametri presenti del modello ristretto sono stimati esattamente come nel modello esteso dal momento che seguono le stesse distribuzioni di probabilità.

Di seguito, si mostra come ottenere le *full conditional probabilities* della *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA* come casi particolari di quelle del modello proposto; inoltre, nel caso dell'*Hashtag-LDA* si spiega il motivo per cui la *full conditional probability* di z^*_{ij} derivata in questa tesi non coincide con quella presentata in F. Zhao et al., 2016.

4.4.1 Latent Dirichlet Allocation

La *full conditional probability* di z^V_{ijk} nella *Latent Dirichlet Allocation* può essere ricavata a partire dalla distribuzione di z^V_{ijk} dati $y^V_{ijk} = 1$, $x_{ij} = 1$ e tutte le altre variabili:

$$p(z^V_{ijk} = t^V | z^V_{ijk} = t^V, w_{ijk} = w^*, \ldots) \propto \frac{\alpha_0 + \lambda_{ij,t^V}\alpha_{t^V} + n^{t^V}_{z^V_{ij,-k},z^H_{ij}}}{T\alpha_0 + \sum_{t=1}^{T} \lambda_{ij,t}\alpha_t + N_{ij} + L_{ij} - 1}$$

$$\times \frac{\beta_v^V + n_{y^V=1,-ijk}^{w^*,t^V}}{\sum_{v=1}^V \beta_v^V + n_{y^V=1,-ijk}^{\cdot,t^V}}$$

Il modello non considera gli *hashtag*, quindi tutte le quantità legate ad essi sono nulle:

$$\propto \frac{\alpha_0 + \lambda_{ij,t^V}\alpha_{t^V} + n_{z_{ij,-k}^V}^{t^V}}{T\alpha_0 + \sum_{t=1}^T \lambda_{ij,t}\alpha_t + N_{ij} - 1} \times \frac{\beta_v^V + n_{y^V=1,-ijk}^{w^*,t^V}}{\sum_{v=1}^V \beta_v^V + n_{y^V=1,-ijk}^{\cdot,t^V}}$$

Avendo $y_{udn}^V = 1$ per ogni parola udn e $\lambda_{ij,t} = 1$ per ogni topic t in ogni documento ud, si ottiene:

$$\propto \frac{\alpha_0 + \alpha_{t^V} + n_{z_{ij,-k}^V}^{t^V}}{T\alpha_0 + \sum_{t=1}^T \alpha_t + N_{ij} - 1} \times \frac{\beta_v^V + n_{-ijk}^{w^*,t^V}}{\sum_{v=1}^V \beta_v^V + n_{-ijk}^{\cdot,t^V}}$$

Infine, fissando $\alpha = \alpha_0 + \alpha_1 = \ldots = \alpha_0 + \alpha_T$ e $\beta^V = \beta_1^V = \ldots = \beta_V^V$, si ottiene la *full conditional probability* per effettuare l'inferenza della *Latent Dirichlet Allocation* introdotta in Griffiths e Steyvers, 2004.

4.4.2 Twitter-LDA

La *full conditional probability* di z_{ij}^* in *Twitter-LDA* può essere ricavata a partire dalla distribuzione di z_{ij}^* dati $x_{ij} = 0$ e tutte le altre variabili:

$$p(z_{ij}^* = t^* | \ldots) \propto \frac{\alpha_{t^*}^* + n_{z_i^*,-ij}^{t^*}}{\sum_{t=1}^T \alpha_t^* + D_i - 1} \times \frac{\prod_{v=1}^V \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y_{ijn}^V=1} - 1} (\beta_v^V + n_{y^V=1,-ij}^{v,t^*} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y_{ijn}^V=1} - 1} (\sum_{v=1}^V \beta_v^V + n_{y^V=1,-ij}^{\cdot,t^*} + q)} \times \frac{\prod_{h=1}^H \prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{h_{ijl}=h} \mathbb{1}_{y_{ijl}^H=1} - 1} (\beta_h^H + n_{y^H=1,-ij}^{h,t^*} + q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}} \mathbb{1}_{y_{ijl}^H=1} - 1} (\sum_{h=1}^H \beta_h^H + n_{y^H=1,-ij}^{\cdot,t^*} + q)}$$

Il modello non considera gli *hashtag*, quindi tutte le quantità legate ad essi sono nulle:

$$= \frac{\alpha^*_{t^*} + n^{t^*}_{z^*_i,-ij}}{\sum_{t=1}^{T} \alpha^*_t + D_i - 1} \times \frac{\prod_{v=1}^{V} \prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{w_{ijn}=v} \mathbb{1}_{y^V_{ijn}=1} - 1} (\beta^V_v + n^{v,t^*}_{y^V=1,-ij} + q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}} \mathbb{1}_{y^V_{ijn}=1} - 1} (\sum_{v=1}^{V} \beta^V_v + n^{\cdot,t^*}_{y^V=1,-ij} + q)}$$

La *full conditional probability* di y^V_{ijk} in *Twitter-LDA* coincide con la distribuzione di y^V_{ijk} dati $x_{ij} = 0$ e tutte le altre variabili:

$$p(y^V_{ijk} = 1 | z^*_{ij} = t^*, w_{ijk} = w^*, \ldots) \propto \frac{b^V_1 + n^1_{y^V,-ijk}}{b^V_1 + b^V_2 + N - 1} \times \frac{\beta^V_{w^*} + n^{w^*,t^*}_{y^V=1,-ijk}}{\sum_{v=1}^{V} \beta^V_v + n^{\cdot,t^*}_{y^V=1,-ijk}}$$

$$p(y^V_{ijk} = 0 | w_{ijk} = w^*, \ldots) \propto \frac{b^V_2 + n^0_{y^V,-ijk}}{b^V_1 + b^V_2 + N - 1} \times \frac{\beta^V_{w^*} + n^{w^*}_{y^V=0,-ijk}}{\sum_{v=1}^{V} \beta^V_v + n^{\cdot}_{y^V=0,-ijk}}$$

Infine, fissando $\alpha^* = \alpha^*_1 = \ldots = \alpha^*_T$, $\beta^V = \beta^V_1 = \ldots = \beta^V_V$ e $b^V = b^V_1 = b^V_2$, si ottengono le *full conditional probabilities* per effettuare l'inferenza di *Twitter-LDA*.

In realtà, W. X. Zhao et al., 2011 non sviluppano la *full conditional probability* di z^*_{ij} e considerano una distribuzione condizionata con ancora le funzioni Gamma al suo interno: dal momento che Resnik e Hardisty, 2010 consigliano di rimuoverle, si preferisce utilizzare la formulazione senza funzioni Gamma ottenuta come caso particolare del modello proposto.

4.4.3 Hashtag-LDA

La *full conditional probability* di z^*_{ij} in *Hashtag-LDA* può essere ricavata a partire dalla distribuzione di z^*_{ij} dati $x_{ij} = 0$ e tutte le altre variabili:

$$p(z^*_{ij} = t^* | \ldots) \propto \frac{\alpha^*_{t^*} + n^{t^*}_{z^*_i,-ij}}{\sum_{t=1}^{T} \alpha^*_t + D_i - 1}$$

$$\times \frac{\prod_{v=1}^{V}\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{w_{ijn}=v}\mathbb{1}_{y^V_{ijn}=1}-1}(\beta_v^V + n^{v,t^*}_{y^V=1,-ij}+q)}{\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{y^V_{ijn}=1}-1}(\sum_{v=1}^{V}\beta_v^V + n^{\cdot,t^*}_{y^V=1,-ij}+q)}$$
$$\times \frac{\prod_{h=1}^{H}\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}-1}(\beta_h^H + n^{h,t^*}_{y^H=1,-ij}+q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{y^H_{ijl}=1}-1}(\sum_{h=1}^{H}\beta_h^H + n^{\cdot,t^*}_{y^H=1,-ij}+q)}$$

Avendo $y^V_{udn}=1$ per ogni parola udn, si ottiene:

$$\frac{\alpha^*_{t^*}+n^{t^*}_{z^*_i,-ij}}{\sum_{t=1}^{T}\alpha^*_t + D_i - 1}$$
$$\times \frac{\prod_{v=1}^{V}\prod_{q=0}^{\sum_{n=1}^{N_{ij}}\mathbb{1}_{w_{ijn}=v}-1}(\beta_v^V + n^{v,t^*}_{-ij}+q)}{\prod_{q=0}^{N_{ij}-1}(\sum_{v=1}^{V}\beta_v^V + n^{\cdot,t^*}_{-ij}+q)}$$
$$\times \frac{\prod_{h=1}^{H}\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{h_{ijl}=h}\mathbb{1}_{y^H_{ijl}=1}-1}(\beta_h^H + n^{h,t^*}_{y^H=1,-ij}+q)}{\prod_{q=0}^{\sum_{l=1}^{L_{ij}}\mathbb{1}_{y^H_{ijl}=1}-1}(\sum_{h=1}^{H}\beta_h^H + n^{\cdot,t^*}_{y^H=1,-ij}+q)}$$

La *full conditional probability* di y^H_{ijk} in *Twitter-LDA* coincide con la distribuzione di y^H_{ijk} dati $x_{ij}=0$ e tutte le altre variabili:

$$p(y^H_{ijk}=1|z^*_{ij}=t^*,h_{ijk}=h^*,\ldots) \propto \frac{b_1^H + n^1_{y^H,-ijk}}{b_1^H+b_2^H+L-1} \times \frac{\beta^H_{h^*}+n^{h^*,t^*}_{y^H=1,-ijk}}{\sum_{h=1}^{H}\beta_h^H + n^{\cdot,t^*}_{y^H=1,-ijk}}$$
$$p(y^H_{ijk}=0|h_{ijk}=h^*,\ldots) \propto \frac{b_2^H + n^0_{y^H,-ijk}}{b_1^H+b_2^H+L-1} \times \frac{\beta^H_{h^*}+n^{h^*}_{y^H=0,-ijk}}{\sum_{h=1}^{H}\beta_h^H + n^{\cdot}_{y^H=0,-ijk}}$$

Infine, fissando $\alpha^* = \alpha^*_1 = \ldots = \alpha^*_T$, $\beta^V = \beta^V_1 = \ldots = \beta^V_V$, $\beta^H = \beta^H_1 = \ldots = \beta^H_H$ e $b^H = b^H_1 = b^H_2$, si ottengono le *full conditional probabilities* per effettuare l'inferenza di *Hashtag-LDA*.

Capitolo 5

Implementazione

Fin'ora si è posta l'attenzione sugli aspetti teorici del *Collapsed Gibbs Sampler* senza entrare nello specifico su come esso viene effettivamente implementato: in questo capitolo si sposta attenzione sulle accortezze necessarie per ottenere un algoritmo funzionante che permetta di ottenere delle stime attendibili in un periodo di tempo non eccessivamente lungo.

Per implementare il *Collapsed Gibbs Sampler* si riprende l'idea alla base dell'implementazione dell'algoritmo proposto in Griffiths e Steyvers, 2004 per l'inferenza della *Latent Dirichlet Allocation* e la si adatta al nuovo caso più complesso. Le *full conditional probabilities* derivate in sottosezione 4.2.4 contengono dei conteggi calcolati sull'intera collezione di documenti: calcolare queste quantità sarebbe estremamente oneroso computazionalmente dal momento che sarebbe necessario scorrere l'intera collezione ogni volta che deve essere calcolata la *full conditional probability* di una variabile latente. Una soluzione efficiente è definire delle matrici di conteggi che vengono aggiornate man mano che l'algoritmo campiona nuovi valori per le variabili latenti del *topic model*. Così facendo è sì necessario avere più elementi salvati in memoria contemporaneamente, sia l'ultimo stato della catena di Markov sia le matrici di conteggi, tuttavia si ottiene un algoritmo estremamente più efficiente dal momento che è più veloce aggiornare le matrici di conteggi ogni volta che lo stato della catena viene modificato piuttosto che calcolare ogni volta i conteggi da zero.

L'implementazione è riportata in Algoritmo 2: in questo caso, oltre a inizializzare il primo stato della catena, è necessario inizializzare anche le

matrici di conteggi; una volta fissate queste quantità, inizia il processo di esplorazione dello spazio degli stati tramite l'aggiornamento delle variabili latenti. Si noti che in memoria si ha unico stato che viene continuamente aggiornato: per tenere traccia degli stati assunti dalla catena alla fine di ogni iterazione dell'algoritmo, questi vengono salvati in formato `.rds`[1].

Algoritmo 2 *Collapsed Gibbs Sampler*

Input: parole $\mathbf{w}$, *hashtag* $\mathbf{h}$, utenti $\mathbf{u}$, parametri $\boldsymbol{\alpha}^*$, $\boldsymbol{\alpha}$, α_0, $\boldsymbol{\beta}^V$, $\boldsymbol{\beta}^H$, $\mathbf{b}$

Inizializzazione dello stato iniziale e delle matrici di conteggi
for $i = 1, \ldots, I$ **do**
    for $ud = 1, \ldots, D$ **do**
        Aggiornamento di x_{ud} e dei conteggi
        Aggiornamento di z^*_{ud} e dei conteggi
        for $t = 1, \ldots, T$ **do**
            Aggiornamento di $\lambda_{ud,t}$ e dei conteggi
        end for
        for $n = 1, \ldots, N_{ud}$ **do**
            Aggiornamento di y^V_{udn} e dei conteggi
            Aggiornamento di z^V_{udn} e dei conteggi
        end for
        for $l = 1, \ldots, L_{ud}$ **do**
            Aggiornamento di y^H_{udl} e dei conteggi
            Aggiornamento di z^H_{udl} e dei conteggi
        end for
    end for
    Salvataggio in formato `.rds` dello stato i
end for

Il linguaggio utilizzato per l'implementazione è *R* poiché, a discapito dell'efficienza, permette di costruire l'algoritmo in maniera relativamente semplice; in Appendice C si riporta il codice. Nella prossime sezioni si analizzano i seguenti aspetti riguardanti l'algoritmo sopra esposto:

- definizione delle matrici in cui sono salvati le variabili latenti $\mathbf{z}^V$, $\mathbf{y}^V$, $\mathbf{z}^H$, $\mathbf{y}^H$, $\mathbf{x}$, $\mathbf{z}^*$ e $\boldsymbol{\lambda}$;
- definizione delle matrici di conteggi;
- inizializzazione dello stato iniziale e delle matrici di conteggi;

[1] `.rds` è un formato nativo di *R* per il salvataggio di una singola variabile in un *file*.

- aggiornamento delle variabili latenti e dei conteggi ad esse legati.

Infine, nell'ultima sezione si mostra come ottenere le stime *Monte Carlo* delle variabili latenti e dei parametri a partire dalla catena di Markov.

5.1 Matrici dello Stato della Catena di Markov

La variabili latenti di cui si vuole ottenere una stima a posteriori –$\mathbf{z}^V$, $\mathbf{y}^V$, $\mathbf{z}^H$, $\mathbf{y}^H$, $\mathbf{x}$, $\mathbf{z}^*$ e $\boldsymbol{\lambda}$– sono tutte a livello di documento, a livello di parola o a livello di *hashtag*, si opta quindi per considerare gli indici *ud* come un unico valore che identifica un documento nella collezione. Così facendo, si rende più semplice l'implementazione dal momento che si riduce di uno il numero di indici di ogni quantità, rendendo più semplice la definizione di strutture per i dati in *R*. Inoltre, si introduce un vettore $\mathbf{u}$ di dimensione $D \times 1$ il cui *ud*-mo elemento indica l'utente che ha scritto il *ud*-mo documento nella collezione.[2]

Le variabili a livello di documento sono quindi salvate come vettori: $\mathbf{x}$ e $\mathbf{z}^*$ sono dei vettori $D \times 1$, mentre $\boldsymbol{\lambda}$ è una matrice $D \times T$ così definita

$$\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_D \end{bmatrix} = \begin{bmatrix} \lambda_{1,1} & \dots & \lambda_{1,T} \\ \vdots & \ddots & \vdots \\ \lambda_{D,1} & \dots & \lambda_{D,T} \end{bmatrix}$$

Le variabili a livello di parola sono salvate in matrici $D \times N_{max}$, dove N_{max} è il numero massimo di parole osservate in un documento della collezione:

$$\mathbf{y}^V = \begin{bmatrix} \mathbf{y}_1^V \\ \vdots \\ \mathbf{y}_D^V \end{bmatrix} = \begin{bmatrix} \mathbf{y}_{1,1}^V & \dots & \mathbf{y}_{1,N_{max}}^V \\ \vdots & \ddots & \vdots \\ \mathbf{y}_{D,1}^V & \dots & \mathbf{y}_{D,N_{max}}^V \end{bmatrix}$$

$$\mathbf{z}^V = \begin{bmatrix} \mathbf{z}_1^V \\ \vdots \\ \mathbf{z}_D^V \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{1,1}^V & \dots & \mathbf{z}_{1,N_{max}}^V \\ \vdots & \ddots & \vdots \\ \mathbf{z}_{D,1}^V & \dots & \mathbf{z}_{D,N_{max}}^V \end{bmatrix}$$

[2]Si continua a scrivere *ud*, ma questo indice è un unico valore in {1,...,D}.

Analogamente, le variabili a livello di *hashtag* sono salvate in matrici $D \times L_{max}$, dove L_{max} è il numero massimo di *hashtag* osservati in un documento della collezione:

$$\mathbf{y}^H = \begin{bmatrix} \mathbf{y}_1^H \\ \vdots \\ \mathbf{y}_D^H \end{bmatrix} = \begin{bmatrix} \mathbf{y}_{1,1}^H & \cdots & \mathbf{y}_{1,L_{max}}^H \\ \vdots & \ddots & \vdots \\ \mathbf{y}_{D,1}^H & \cdots & \mathbf{y}_{D,L_{max}}^H \end{bmatrix}$$

$$\mathbf{z}^H = \begin{bmatrix} \mathbf{z}_1^H \\ \vdots \\ \mathbf{z}_D^H \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{1,1}^H & \cdots & \mathbf{z}_{1,L_{max}}^H \\ \vdots & \ddots & \vdots \\ \mathbf{z}_{D,1}^H & \cdots & \mathbf{z}_{D,L_{max}}^H \end{bmatrix}$$

Anche $\mathbf{w}$ e $\mathbf{h}$ sono salvate come matrici $D \times N_{max}$ e $D \times L_{max}$; infine, si noti che queste ultime sei matrici sono sparse poiché elementi che corrispondono a termini –parole e *hashtag*– non esistenti sono fissati a zero.

5.2 Matrici di Conteggi

I conteggi contenuti nelle *full conditional probabilities* possono essere unici per l'intera collezione oppure essere calcolati per ogni topic, ogni parola, ogni *hashtag* o combinazione di essi. Nel secondo caso, è ragionevole raggruppare i conteggi dello stesso tipo in vettori o matrici:

1. Il numero di documenti con più topic $n_{x_u}^1 = \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=1}$ e il numero di documenti con un unico topic $n_{x_u}^0 = \sum_{d=1}^{D_u} \mathbb{1}_{x_{ud}=0}$ scritti dagli U utenti sono raggruppati in una matrice $D \times 2$ definita come

$$n_x = \begin{bmatrix} n_{x_1}^1 & n_{x_1}^0 \\ \vdots & \vdots \\ n_{x_U}^1 & n_{x_U}^0 \end{bmatrix}$$

   Nell'implementazione del *Collapsed Gibbs Sampler* solo la prima colonna viene salvata in memoria poiché i conteggi contenuti nella seconda possono essere ricavati come $n_{x_u}^0 = D_u - n_{x_u}^1$, dove D_u è il numero di documenti scritti dall'utente u.
2. I conteggi che indicano il numero di volte la parola v è associata al topic t considerando sia il tipo di ogni documento sia l'origine di ogni parola, $n_{y^V=1}^{v,t} = \sum_{udn} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^V=1} (\mathbb{1}_{z_{udn}^V=t} \mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z_{ud}^*=t} \mathbb{1}_{x_{ud}=0})$,

sono raggruppati in una matrice $V \times T$ definita come

$$n_{y^V=1}^{1:V,1:T} = \begin{bmatrix} n_{y^V=1}^{1:V,1} & \dots & n_{y^V=1}^{1:V,T} \end{bmatrix} = \begin{bmatrix} n_{y^V=1}^{1,1} & \dots & n_{y^V=1}^{1,T} \\ \vdots & \ddots & \vdots \\ n_{y^V=1}^{V,1} & \dots & n_{y^V=1}^{V,T} \end{bmatrix}$$

La somma $n_{y^V=1}^{\cdot,t}$ degli elementi della t-ma colonna, indicata con $n_{y^V=1}^{1:V,t}$, è pari al numero totale di parole associate al topic t nella collezione considerando sia il tipo di ogni documento sia l'origine di ogni parola:

$$n_{y^V=1}^{\cdot,t} = \sum_{v=1}^{V} n_{y^V=1}^{v,t} = \sum_{udn} \mathbb{1}_{y_{udn}^V=1}(\mathbb{1}_{z_{udn}^V=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z_{ud}^*}\mathbb{1}_{x_{ud}=0})$$

3. I conteggi che indicano il numero di volte l'*hashtag* h è associato al topic t considerando sia il tipo di ogni documento sia l'origine di ogni *hashtag*, $n_{y^H=1}^{h,t} = \sum_{udl} \mathbb{1}_{h_{udl}=h}\mathbb{1}_{y_{udl}^H=1}(\mathbb{1}_{z_{udl}^H=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z_{ud}^*=t}\mathbb{1}_{x_{ud}=0})$, sono raggruppati in una matrice $H \times T$ definita come

$$n_{y^H=1}^{1:H,1:T} = \begin{bmatrix} n_{y^H=1}^{1:H,1} & \dots & n_{y^H=1}^{1:H,T} \end{bmatrix} = \begin{bmatrix} n_{y^H=1}^{1,1} & \dots & n_{y^H=1}^{1,T} \\ \vdots & \ddots & \vdots \\ n_{y^H=1}^{H,1} & \dots & n_{y^H=1}^{H,T} \end{bmatrix}$$

La somma $n_{y^H=1}^{\cdot,t}$ degli elementi della t-ma colonna, indicata con $n_{y^H=1}^{1:H,t}$, è pari al numero totale di *hashtag* associati al topic t nella collezione considerando sia il tipo di ogni documento sia l'origine di ogni *hashtag*:

$$n_{y^H=1}^{\cdot,t} = \sum_{h=1}^{H} n_{y^H=1}^{h,t} = \sum_{udl} \mathbb{1}_{y_{udl}^H=1}(\mathbb{1}_{z_{udl}^H=t}\mathbb{1}_{x_{ud}=1} + \mathbb{1}_{z_{ud}^*}\mathbb{1}_{x_{ud}=0})$$

4. I conteggi che indicano il numero di documenti scritti dall'utente u il cui topic principale è t, $n_{z_u^*}^t = \sum_{d=1}^{D_u} \mathbb{1}_{z_{ud}^*=t}$, sono raggruppati in una matrice $U \times T$ definita come

$$n_{z^*} = \begin{bmatrix} n_{z_1^*}^{1:T} \\ \vdots \\ n_{z_U^*}^{1:T} \end{bmatrix} = \begin{bmatrix} n_{z_1^*}^{1} & \dots & n_{z_1^*}^{T} \\ \vdots & \ddots & \vdots \\ n_{z_U^*}^{1} & \dots & n_{z_U^*}^{T} \end{bmatrix}$$

Si noti che vale $D_u = \sum_{t=1}^{T} n_{z_u^*}^t$, ovvero la somma degli elementi della u-ma riga è pari al numero totale di documenti scritti dall'utente u.

5. Il numero totale di topic attivi $n_{\lambda}^{1} = \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=1}$ e non attivi $n_{\lambda}^{0} = \sum_{udt} \mathbb{1}_{\lambda_{ud,t}=0}$ nella collezione sono raggruppati in un vettore 2×1 definito come
$$n_{\lambda} = \begin{bmatrix} n_{\lambda}^{1} & n_{\lambda}^{0} \end{bmatrix}$$
Nell'implementazione del *Collapsed Gibbs Sampler* solo n_{λ}^{1} viene salvato in memoria poiché il secondo conteggio può essere ricavato come $n_{\lambda}^{0} = DT - n_{\lambda}^{1}$, dove DT è il numero totale di volte in cui si valuta se un topic è attivo o meno nei un documenti della collezione.
6. Il numero di parole generate a partire da un topic $n_{y^V}^{1} = \sum_{udn} \mathbb{1}_{y_{udn}^{V}=1}$ e il numero di parole di sottofondo $n_{y^V}^{0} = \sum_{udn} \mathbb{1}_{y_{udn}^{V}=0}$ nella collezione sono raggruppati in un vettore 2×1 definito come
$$n_{y^V} = \begin{bmatrix} n_{y^V}^{1} & n_{y^V}^{0} \end{bmatrix}$$
Anche in questo caso il secondo conteggio non viene salvato in memoria poiché può essere ricavato come $n_{y^V}^{0} = N - n_{y^V}^{1}$, dove N è il numero totale di parole nella collezione.
7. I conteggi che indicano il numero di volte che v appare come parola di sottofondo nella collezione, $n_{y^V=0}^{v} = \sum_{udn} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y_{udn}^{V}=0}$, sono raggruppati in un vettore $V \times 1$ definito come
$$n_{y^V=0}^{1:V} = \begin{bmatrix} n_{y^V=0}^{1} \\ \vdots \\ n_{y^V=0}^{V} \end{bmatrix}$$
La somma degli elementi del vettore, indicata con $n_{y^V=0}^{\cdot}$, è pari al numero totale di parole di sottofondo nella collezione:
$$n_{y^V=0}^{\cdot} = \sum_{v=1}^{V} n_{y^V=0}^{v} = \sum_{udn} \mathbb{1}_{y_{udn}^{V}=0}$$
8. Il numero di *hashtag* generati a partire da un topic $n_{y^H}^{1} = \sum_{udl} \mathbb{1}_{y_{udl}^{H}=1}$ e il numero di *hashtag* globali $n_{y^H}^{0} = \sum_{udl} \mathbb{1}_{y_{udl}^{H}=0}$ nella collezione sono raggruppati in un vettore 2×1 definito come
$$n_{y^H} = \begin{bmatrix} n_{y^H}^{1} & n_{y^H}^{0} \end{bmatrix}$$
Anche in questo caso il secondo conteggio non viene salvato in memoria poiché può essere ricavato come $n_{y^H}^{0} = L - n_{y^H}^{1}$, dove L è il numero totale di *hashtag* nella collezione.

9. I conteggi che indicano il numero di volte che h appare come *hashtag globale*, $n^{h}_{y^H=0} = \sum_{udl} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=0}$, sono raggruppati in un vettore $H \times 1$ definito come

$$n^{1:H}_{y^H=0} = \begin{bmatrix} n^{1}_{y^H=0} \\ \vdots \\ n^{H}_{y^H=0} \end{bmatrix}$$

La somma degli elementi del vettore, indicata con $n^{\cdot}_{y^H=0}$, è pari al numero totale di *hashtag globali* nella collezione:

$$n^{\cdot}_{y^H=0} = \sum_{h=1}^{H} n^{h}_{y^H=0} = \sum_{udl} \mathbb{1}_{y^H_{udl}=0}$$

10. I conteggi che indicano il numero di elementi testuali –parole, *hashtag* o entrambi– del documento ud a cui è associato il topic t senza considerare né il tipo di ogni documento né l'origine di ogni elemento, $n^{t}_{z^V_{ud}, z^H_{ud}} = \sum_{n=1}^{N_{ud}} \mathbb{1}_{z^V_{udn}=t} + \sum_{l=1}^{L_{ud}} \mathbb{1}_{z^H_{udl}=t}$, sono raggruppati in una matrice $D \times T$ definita come

$$n^{1:T}_{z^V, z^H} = \begin{bmatrix} n^{1:T}_{z^V_{11}, z^H_{11}} \\ \vdots \\ n^{1:T}_{z^V_{UD_U}, z^H_{UD_U}} \end{bmatrix} = \begin{bmatrix} n^{1}_{z^V_{11}, z^H_{11}} & \cdots & n^{T}_{z^V_{11}, z^H_{11}} \\ \vdots & \ddots & \vdots \\ n^{1}_{z^V_{UD_U}, z^H_{UD_U}} & \cdots & n^{T}_{z^V_{UD_U}, z^H_{UD_U}} \end{bmatrix}$$

Si noti che vale $N_{ud} + L_{ud} = \sum_{t=1}^{T} n^{t}_{z^V_{ud}, z^H_{ud}}$, ovvero la somma degli elementi della ud-ma riga è pari al numero totale elementi testuali del d-mo documento dell'utente u.

Nelle *full conditional probabilities* di x_{ud} e z^*_{ud} sono presenti due ulteriori conteggi relativi alle parole contenute in ogni documento:

$$n^{v,t}_{z^V_{ud}, y^V=1} = \sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{z^V_{udn}=t}$$

$$n^{v}_{y^V_{ud}=1} = \sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1}$$

e due ulteriori conteggi relativi agli *hashtag* contenuti in ogni documento:

$$n^{h,t}_{z^H_{ud}, y^H=1} = \sum_{l=1}^{L_{ud}} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1} \mathbb{1}_{z^H_{udl}=t}$$

$$n^{h}_{y^H_{ud}=1} = \sum_{l=1}^{L_{ud}} \mathbb{1}_{h_{udl}=h} \mathbb{1}_{y^H_{udl}=1}$$

Ricordando che a ogni parola può essere associato un unico topic, ovvero vale $\sum_{t=1}^{T} \mathbb{1}_{z^V_{udn}=t} = 1$ per ogni parola udn, il secondo conteggio può essere ottenuto a partire dal primo:

$$\begin{aligned} n^{v}_{y^V_{ud}=1} &= \sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \\ &= \sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \sum_{t=1}^{T} \mathbb{1}_{z^V_{udn}=t} \\ &= \sum_{t=1}^{T} \sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v} \mathbb{1}_{y^V_{udn}=1} \mathbb{1}_{z^V_{udn}=t} \\ &= \sum_{t=1}^{T} n^{v,t}_{z^V_{ud},y^V=1} \end{aligned}$$

Analogamente, ricordando che a ogni *hashtag* può essere associato un unico topic, ovvero vale $\sum_{t=1}^{T} \mathbb{1}_{z^H_{udl}=t} = 1$ per ogni *hashtag* udl, il quarto conteggio può essere ottenuto a partire dal terzo:

$$n^{h}_{y^H_{ud}=1} = \sum_{t=1}^{T} n^{h,t}_{z^H_{ud},y^H=1}$$

Quindi, nell'implementazione è ragionevole salvare in memoria solo le matrici di conteggi relative a $n^{v,t}_{z^V_{ud},y^V=1}$ e $n^{h,t}_{z^H_{ud},y^H=1}$, e ricavare le due rimanenti a partire dalle prime.

I conteggi che indicano il numero di volte la parola v non di sottofondo è associata al topic t nel documento ud, $n^{v,t}_{z^V_{ud},y^V=1}$, sono raggruppati in una lista di D matrici $V \times T$ definite come

$$n^{1:V,1:T}_{z^V_{ud},y^V=1} = \begin{bmatrix} n^{1:V,1}_{y^V=1} & \cdots & n^{1:V,T}_{y^V=1} \end{bmatrix} = \begin{bmatrix} n^{1,1}_{z^V_{ud},y^V=1} & \cdots & n^{1,T}_{z^V_{ud},y^V=1} \\ \vdots & \ddots & \vdots \\ n^{V,1}_{z^V_{ud},y^V=1} & \cdots & n^{V,T}_{z^V_{ud},y^V=1} \end{bmatrix}$$

dove $ud = 1, \ldots, D$. Il conteggio che indica il numero di volte la parola v è generata a partire da un topic nel documento ud, $n^{v}_{y^V_{ud}=1}$, è ottenuto come la somma degli elementi della v-ma riga di $n^{1:V,1:T}_{z^V_{ud},y^V=1}$.

I conteggi che indicano il numero di volte l'*hashtag* h non *globale* è associato al topic t nel documento ud, $n^{h,t}_{z^H_{ud},y^H=1}$, sono raggruppati in una lista di D matrici $H \times T$ definite come

$$n^{1:H,1:T}_{z^V_{ud},y^H=1} = \begin{bmatrix} n^{1:H,1}_{y^H=1} & \dots & n^{1:H,T}_{y^H=1} \end{bmatrix} = \begin{bmatrix} n^{1,1}_{z^H_{ud},y^H=1} & \dots & n^{1,T}_{z^H_{ud},y^H=1} \\ \vdots & \ddots & \vdots \\ n^{V,1}_{z^H_{ud},y^H=1} & \dots & n^{H,T}_{z^H_{ud},y^H=1} \end{bmatrix}$$

dove $ud = 1, \dots, D$. Il conteggio che indica il numero di volte l'*hashtag* h è generato a partire da un topic nel documento ud, $n^h_{y^H_{ud}=1}$, è ottenuto come la somma degli elementi della h-ma riga di $n^{1:H,1:T}_{z^H_{ud},y^H=1}$.

Riepilogando, in Tabella 5.1 si riportano per ogni conteggio la dimensione della matrice in cui è contenuto, la lista delle variabili da cui dipende e la lista delle *full conditional probabilities* in cui compare. Si noti che molte di queste matrici sono sparse.

conteggio	dimensione	variabili da cui dipende					distr. in cui compare			
$n^1_{x_u}$	$U \times 1$	$\mathbf{x}_u$					$\mathbf{x}_u$			
$n^{v,t}_{y^V=1}$	$V \times T$	$\mathbf{w}$	$\mathbf{x}$	$\mathbf{z}^*$	$\mathbf{z}^V$	$\mathbf{y}^V$	$\mathbf{x}$	$\mathbf{z}^*$	$\mathbf{z}^V$	$\mathbf{y}^V$
$n^{h,t}_{y^H=1}$	$H \times T$	$\mathbf{h}$	$\mathbf{x}$	$\mathbf{z}^*$	$\mathbf{z}^H$	$\mathbf{y}^H$	$\mathbf{x}$	$\mathbf{z}^*$	$\mathbf{z}^H$	$\mathbf{y}^H$
$n^t_{z^*_u}$	$U \times T$	$\mathbf{z}^*_u$					$\mathbf{z}^*_u$			
n^1_λ	1×1	$\boldsymbol{\lambda}$					$\boldsymbol{\lambda}$			
$n^1_{y^V}$	1×1	$\mathbf{y}^V$					$\mathbf{y}^V$			
$n^v_{y^V=0}$	$V \times 1$	$\mathbf{w}$	$\mathbf{y}^V$				$\mathbf{y}^V$			
$n^1_{y^H}$	1×1	$\mathbf{y}^H$					$\mathbf{y}^H$			
$n^h_{y^H=0}$	$H \times 1$	$\mathbf{h}$	$\mathbf{y}^H$				$\mathbf{y}^H$			
$n^t_{z^V_{ud},z^H_{ud}}$	$D \times T$	$\mathbf{z}^V_{ud}$	$\mathbf{z}^H_{ud}$				$\mathbf{z}^V_{ud}$	$\mathbf{z}^H_{ud}$		
$n^{v,t}_{z^V_{ud},y^V=1}$	$D \times V \times T$	$\mathbf{w}_{ud}$	$\mathbf{z}^V_{ud}$	$\mathbf{y}^V_{ud}$			x_{ud}	z^*_{ud}		
$n^{h,t}_{z^H_{ud},y^H=1}$	$D \times H \times T$	$\mathbf{h}_{ud}$	$\mathbf{z}^H_{ud}$	$\mathbf{y}^H_{ud}$			x_{ud}	z^*_{ud}		

Totale conteggi: $3 + U + V + H + T(V + H + U + D + DV + DH)$

Tabella 5.1: Tabella dei conteggi; in basso si riporta il numero totale di conteggi contenuti nei vettori e nelle matrici.

5.3 Inizializzazione dello Stato Iniziale e delle Matrici di Conteggi

Nel contesto dei *modelli generativi probabilistici*, è possibile sfruttare la rappresentazione del modello come *processo generativo* per inizializzare lo stato iniziale della catena di Markov; in particolare, si segue il processo generativo omettendo i passi in cui sono generati i parametri **par** e quelli in cui si generano le variabili osservate. Più nello specifico, si generano le variabili latenti dalle loro distribuzioni di probabilità, ma sostituendo i parametri su cui si vuole fare inferenza a posteriori con il valore atteso della loro distribuzione a priori. Ad esempio, nel *processo generativo* si ha $x_{ud} \sim Bern(\pi_u^T)$, mentre nell'inizializzazione si ha $x_{ud} \sim Bern\left(\frac{b_1^T}{b_1^T + b_2^T}\right)$, dove $\frac{b_1^T}{b_1^T + b_2^T}$ è il valore atteso di una distribuzione Beta con parametri b_1^T e b_2^T. Così facendo le variabili sono generate casualmente, ma tendono ad assumere i valori che a priori si assumono più probabili. Per semplicità, i topic associati alle parole e agli *hashtag* sono generati da una distribuzione categoriale con parametro $\frac{\boldsymbol{\alpha}}{\sum_{t=1}^T \alpha_t}$ al posto di $\frac{\boldsymbol{\alpha}^{(ud)}}{\sum_{t=1}^T \alpha_t^{(ud)}}$. Infine, man mano che le variabili latenti vengono generate, si effettua anche l'aggiornamento dei conteggi in modo da non dover iterare nuovamente sull'intera collezione in un momento successivo per il calcolo di quest'ultimi. Il procedimento appena esposto è riportato in Algoritmo 3.

5.4 Aggiornamento delle Variabili

In memoria si hanno le matrici di conteggi totali, ovvero che considerano tutte le variabili disponibili, tuttavia la *full conditional probability* di una variabile u_k dipende da conteggi parziali in cui u_k non è presente: l'aggiornamento di una variabile u_k deve quindi necessariamente essere preceduto da alcuni passaggi preliminari. Più nello specifico, prima di tutto si rimuove[3] u_k dai conteggi che dipendono da essa –si veda la colonna *variabili da cui dipende* in Tabella 5.1– in modo tale che i conteggi totali in memoria diventino parziali per u_k. Avendo a disposizione i conteggi parziali, si può precedere con il calcolo della *full conditional probability* di u_k: visto che la

[3]La rimozione di u_k avviene seguendo le scomposizioni dei conteggi introdotte per il calcolo delle *full conditional probabilities* in sottosezione 4.2.4.

Algoritmo 3 Inizializzazione dello stato iniziale e delle matrici di conteggi

Creazione di matrici di conteggi vuote
for $ud = 1, \dots, D$ **do**
    Estrazione di x_{ud} da $Bern\left(\frac{b_1^T}{b_1^T + b_2^T}\right)$
    Aggiornamento dei conteggi
    Estrazione di z^*_{ud} da $Cat\left(\frac{\boldsymbol{\alpha}^*}{\sum_{t=1}^T \alpha_t^*}\right)$
    Aggiornamento dei conteggi
    for $t = 1, \dots, T$ **do**
        Estrazione di $\lambda_{ud,t}$ da $Bern\left(\frac{b_1^\delta}{b_1^\delta + b_2^\delta}\right)$
        Aggiornamento dei conteggi
    end for
    for $n = 1, \dots, N_{ud}$ **do**
        Estrazione di y^V_{udn} da $Bern\left(\frac{b_1^V}{b_1^V + b_2^V}\right)$
        Aggiornamento dei conteggi
        Estrazione di z^V_{udn} da $Cat\left(\frac{\boldsymbol{\alpha}}{\sum_{t=1}^T \alpha_t}\right)$
        Aggiornamento dei conteggi
    end for
    for $l = 1, \dots, L_{ud}$ **do**
        Estrazione di y^H_{udl} da $Bern\left(\frac{b_1^H}{b_1^H + b_2^H}\right)$
        Aggiornamento dei conteggi
        Estrazione di z^H_{udl} da $Cat\left(\frac{\boldsymbol{\alpha}}{\sum_{t=1}^T \alpha_t}\right)$
        Aggiornamento dei conteggi
    end for
end for

distribuzione di un qualsiasi u_k è una distribuzione categoriale[4], a livello pratico la *full conditional probability* è un vettore di pesi che indica le modalità più verosimili da osservare condizionatamente ai rimanenti valori assunti dalle altre variabili del modello. Il nuovo valore di u_k è quindi estratto da una distribuzioni categoriale il cui vettore di probabilità è la *full conditional probability* normalizzata. Infine, aggiornando i conteggi con il nuovo valore di u_k, si ottengono nuovamente i conteggi totali in memoria. Il procedimento appena esposto è riportato in Algoritmo 4.

[4]Una distribuzione di Bernoulli con parametro π coincide con una distribuzione categoriale con vettore di probabilità $(\pi, 1-\pi)$ e supporto $\{1, 0\}$.

Algoritmo 4 Aggiornamento di una variabile latente u_k

1: Rimozione di u_k dai conteggi
2: Calcolo della *full conditional probability*
3: Normalizzazione della *full conditional probability*
4: Estrazione di u_k usando *full conditional probability* normalizzata
5: Aggiunta di u_k ai conteggi

5.5 Stime delle Variabili Latenti e dei Parametri

Per ogni variabile latente, $\mathbf{z}^V$, $\mathbf{y}^V$, $\mathbf{z}^H$, $\mathbf{y}^H$, $\mathbf{x}$, $\mathbf{z}^*$ e $\boldsymbol{\lambda}$, l'Algoritmo 2 fornisce I *file* in formato `.rds` contenenti gli I stati della catena di Markov. Per ottenere le stime *Monte Carlo* delle variabili indicatrici, $\mathbf{y}^V$, $\mathbf{y}^H$,$\mathbf{x}$ e $\boldsymbol{\lambda}$, basta calcolare la media empirica. Ad esempio, per x_{ud} si ha

$$\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} x_{ud}^{(i)}$$

dove $\mathcal{M}$ è l'insieme degli indici degli stati che vengono effettivamente utilizzati per calcolare le stime *Monte Carlo* e l'apice (i) indica che si sta considerando il valore assunto dalla variabile nell'i-mo stato della catena. Il valore assunto dalla stima *Monte Carlo* è la probabilità che quella variabile sia pari a 1; negli esperimenti si assume che una variabile sia pari a 1 se la sua stima *Monte Carlo* è maggiore di 0.5.

Non potendo calcolare la media delle variabili categoriali $\mathbf{z}^V$, $\mathbf{z}^H$ e $\mathbf{z}^*$, questa è sostituita dal voto di maggioranza.

Infine, per ottenere le stime *Monte Carlo* dei parametri, è necessario effettuare un passaggio aggiuntivo in cui si stimano i parametri di ogni stato della catena utilizzando le formule derivate nella sottosezione 4.2.5; una volta ottenuti questi valori, è possibile a procedere al calcolo delle stime *Monte Carlo* come sopra. Ad esempio, per π^V si ha

$$\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \pi^{V(i)} = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \frac{b_1^V + \sum_{udn} \mathbb{1}_{y_{udn}^{V(i)}=1}}{b_1^V + b_2^V + N}$$

Il procedimento appena esposto è riportato in Algoritmo 5.

Algoritmo 5 Stima delle variabili latenti e dei parametri

1: **for** $i \in \mathcal{M}$ **do**
2: Importazione in memoria dello stato i
3: Calcolo delle stime dei parametri nello stato i
4: **end for**
5: Stime *Monte Carlo* delle variabili latenti e dei parametri

Capitolo 6

Esperimenti

Nel Capitolo 3 e nel Capitolo 4 è stato introdotto un nuovo *topic model* e formulato un *Collapsed Gibbs Sampler* per la stima a posteriori della sua struttura latente; nel Capitolo 5 si è posta l'attenzione sull'implementazione dell'algoritmo; infine, in questo capitolo si propone un possibile approccio per valutare il modello proposto sia quantitativamente sia qualitativamente e confrontarlo con i suoi tre casi particolari introdotti nel Capitolo 3: *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA*.

Nella sezione 6.1 si introduce *Twitter* e si mostra il procedimento utilizzato per scaricare i *tweet* ed ottenere un dataset strutturato su cui applicare i quattro *topic model*; nella sezione 6.2 si effettuano delle analisi esplorative sulla collezione di *tweet*; nella sezione 6.3 si analizza la convergenza delle catene di Markov ottenute con i *Collapsed Gibbs Sampler* e si confrontano i vari *topic model* in termini di *topic coherence*; infine, nella sezione 6.4 si mostra come interpretare la struttura latente del modello proposto.

6.1 Costruzione della Collezione di Tweet

Non trovando in letteratura una collezione adatta su cui sfruttare a pieno le potenzialità del modello proposto in questa tesi, si opta per costruirne una ad-hoc utilizzando le *Twitter API*.

Twitter è una piattaforma di *microblogging* nata nel 2006 che si basa su messaggi brevi, detti *tweet*, la cui lunghezza massima è di 280 caratteri[1]. A

[1]Originariamente fissato a 140, nel 2017 il limite è stato spostato a 280.

differenza di altri *social media* in cui la comunicazione tramite post spesso si basa sull'utilizzo congiunto di testo e immagini o video, i *tweet* di *Twitter* tendono a concentrare tutta la loro informazione esclusivamente nel testo. Ciò rende *Twitter* una fonte ideale da cui estrarre informazione attraverso le tecniche di *text mining*, e in questo caso attraverso *topic model*. Inoltre, gli utenti hanno a disposizione tre metodi per visualizzare contenuti altrui:

1. il meccanismo del *follow*[2];
2. la funzionalità di ricerca;
3. la pagina *Esplora*[3].

I metodi 2 e 3 si basano sull'*hashtag*, uno strumento introdotto da *Twitter* stesso nel 2007, che favorisce la diffusione dell'informazione rendendo i *tweet* più facilmente reperibili.

Concludendo, la presenza di *hashtag* nei *tweet*, ovvero post di *microblog*, rende *Twitter* un ottima fonte da cui derivare collezioni su cui valutare il modello proposto.

6.1.1 Download di Tweet tramite Twitter API

Twitter offre a società, sviluppatori e utenti l'accesso programmatico ai suoi dati attraverso delle *Application Programming Interfaces* (*API*), o *interfacce di programmazione delle applicazioni*: esse permettono a uno sviluppatore –inteso come colui che utilizza le *API*– di interagire con le risorse di *Twitter* attraverso una serie di *endpoint* in maniera relativamente semplice ed intuitiva. Una *API* fornisce essenzialmente un linguaggio per far interagire due sistemi informatici ed ogni *endpoint* può essere visto come un diverso punto di contatto: ognuno di essi ha una sua sintassi e permette di ottenere diverse informazioni. Tra le varie funzionalità fornite dagli *endpoint* si ricordano la gestione programmatica di un profilo personale e delle impostazioni di un account, pubblicazione di *tweet*, accesso a *tweet* pubblici tramite ricerca di parole chiave o per utente, strumenti per la creazione e gestione di campagne pubblicitarie (Twitter, n.d.-a). Due *endpoint* permettono di cercare tra i

[2]Se l'utente A segue l'utente B, significa che A è interessato, almeno in parte, a ciò che pubblica B.

[3]Nella pagina *Esplora Twitter* consiglia all'utente contenuti sulla base dell'attività precedentemente svolta sulla piattaforma.

tweet pubblici presenti nella piattaforma: *recent search* e *full-archive search*. Essi condividono lo stesso design e le stesse funzionalità, tuttavia il primo dà accesso ai *tweet* pubblicati nell'ultima settimana, mentre il secondo non impone vincoli temporali. In entrambi i casi, è possibile filtrare i *tweet* tramite una singola *search query* che può essere costruita con una serie di operatori combinati tra loro tramite logica booleana[4] (Twitter, n.d.-c).

I due *endpoint* e il contenuto delle *query* sono vincolati al *level access* dell'account dello sviluppatore: avendo un account di livello *Elevated*, è stato possibile utilizzare solo l'*endpoint recent search* con *query* lunghe al massimo 512 caratteri. Nella costruzione della collezione di *tweet* è stato necessario tenere conto di queste limitazioni e si è quindi optato per eseguire una serie di *query*. In particolare, si utilizza un approccio in tre fasi: nella prima si scaricano 20,000 *tweet* a tema *covid*; nella seconda si scaricano tutti i *tweet* appartenenti alle conversazioni dei *tweet* identificati nella prima fase; infine, nella terza si crea un'unica collezione contenente sia i *tweet* della prima fase sia quelli della seconda. Così facendo si riesce a ottenere un *corpus* di dimensione contenute che non si limita ad essere una collezione di singoli *tweet* potenzialmente tutti indipendenti tra loro, ma una collezione di *tweet* appartenenti a una determinata lista di conversazioni.

Nella prima fase un *tweet* è considerato a tema *covid* se almeno una delle seguenti parole, o l'*hashtag* corrispondente, è presente all'interno del suo testo:

`sars-cov-2`	`covid19`	`sars-cov2`	`pandemia`
`covid`	`corona virus`	`coronavirus`	

Tenendo conto dell'approccio appena descritto e dei vincoli a cui si è sottoposti, si è seguito il seguente procedimento:

(1) Fase 1

   i. Costruzione della *query*: *tweet* in italiano pubblicati tra le 22:00 del 24/01/2022 e le 22:00 del 30/01/2022 che non sono *retweet* e che contengono almeno una delle parole chiave.
   ii. Download di 20,000 *tweet* che soddisfano la *query* del punto (1.i).

[4]Si rimanda a Twitter, n.d.-b per una trattazione dettagliata su come costruire *query* combinando diversi operatori.

(2) Fase 2

   i. Costruzione delle *query*: *tweet* in italiano che appartengono alle conversazioni dei *tweet* ottenuti al punto (1.ii).
   ii. Download di *tweet* che soddisfano le *query* del punto (2.i).

(3) Fase 3

   i. Creazione di un'unica collezione contenente i *tweet* scaricati ai punti (1.ii) e (2.ii).
   ii. Rimozione dei *tweet* duplicati nella collezione creata la punto (3.i);
   iii. Salvataggio della collezione ottenuta al punto (3.ii) in formato `.csv`.

La *query* al punto (1.i) è ottenuta legando sette operatori `keyword`, un operatore `is:retweet` preceduto dalla logica `NOT` e un operatore `lang:` con logiche `OR` tra operatori delle stesso tipo e logiche `AND` tra operatori di tipo diverso; tenendo a mente che la logica `AND` non dev'essere scritta esplicitamente, la *query* è così definita:

```
(sars-cov-2 OR covid19 OR sars-cov2 OR pandemia OR covid OR
corona virus OR coronavirus) lang:it -is:retweet
```

Si noti che il vincolo sulla data di pubblicazione non è applicato esplicitamente attraverso la *query*, ma tramite la funzione della libreria `tweepy` (Roesslein, n.d.) di *Python* utilizzata per interfacciarsi con le *Twitter API*.

Dato il limite di lunghezza, al punto (2.i) si crea una *query* diversa per ogni 10 conversazioni; una di queste è ottenuta legando dieci operatori `conversation_id:` e un operatore `lang:` con logiche `OR` tra operatori delle stesso tipo e logiche `AND` tra operatori di tipo diverso; la *query* è quindi così definita:

```
(conversation_id:<ID01> OR ... OR conversation_id:<ID10>)
lang:it
```

dove `<ID01>` è l'identificativo della prima conversazione considerata dalla *query*, `<ID02>` quello della seconda e così via. Si noti che l'operatore `lang:` è necessario poiché le conversazioni in *Twitter* non sono necessariamente da solo *tweet* scritti nella stessa lingua.

6.1.2 Pulizia dei Tweet

Per poter utilizzare tecniche di *text mining* –in questo caso i *topic model*– è necessario convertire i testi dei documenti della collezione –dati non strutturati– in un insieme di dati quantitativi. In particolare, l'obiettivo della pulizia dei testi è rappresentare ogni documento come una *bag-of-words*, ovvero un insieme di termini di cui non è rilevante l'ordine; essenzialmente, si vuole ottenere una rappresentazione semplificata in cui è d'interesse solo sapere se un termine è contenuto in un documento e quante volte. Si effettua una pulizia dei testi dei *tweet* percorrendo le seguenti fasi:

1. Tokenizzazione
2. Fase di normalizzazione
    1. Conversione degli *unicode* corrispondenti alle *emoji* in stringhe autoesplicative[5].
    2. Conversione del testo in minuscolo.
    3. Rimozione di punteggiatura, *url* e eventuali stringhe in *html*.
    4. Sostituzione di lettere accentate con la corrispondente versione non accentata.
3. Prima fase di filtraggio
    1. Rimozione delle *stopwords*.
    2. Rimozione delle stringhe utilizzate per menzionare altri utenti[6].
    3. Rimozione di parole formate solo da numeri, *unicode*, ...
4. *Stemming* applicato a tutti i termini eccetto le *emoji* e gli *hashtag*.
5. Seconda fase di filtraggio
    1. Rimozione delle parole formate da un unico carattere.
    2. Rimozione degli *hapax*[7].

Si noti che con termine si fa riferimento a un qualsiasi elemento contenuto nel testo di un *tweet* e che un termine può essere una parola o un *hashtag*; quindi, un'operazione applicata ai termini influenza sia le parole sia gli *hashtag*.

[5] La lista di riferimento per la conversione emote-stringa è riportata in Consortium, n.d.

[6] Un utente può essere menzionato in *Twitter* inserendo il simbolo @ davanti al suo nome utente.

[7] Un *hapax legomenon*, o *hapax*, è una parola che ricorre una sola volta in una collezione di documenti.

6.1.3 Filtro dei Tweet

Attraverso il procedimento introdotto nella sezioni precedenti si ottiene una collezione formata da 204,008 *tweet* con un vocabolario delle parole di ampiezza 24,115 e uno degli *hashtag* di ampiezza 1,969; i *tweet*, suddivisi in 15,470 conversazioni, sono pubblicati da 51,521 utenti distinti.

Data l'elevato tempo necessario per effettuare un numero ragionevole di iterazioni, si filtra ulteriormente la collezione considerando solo i *tweet* pubblicati da utenti attivi nell'intervallo temporale considerato. Più nello specifico, un utente è considerato attivo se la collezione scaricata attraverso le *Twitter API* contiene almeno 70 suoi *tweet*. Infine, riprendendo quanto fatto in W. X. Zhao et al., 2011 e F. Zhao et al., 2016, si effettua un ulteriore filtro con cui vengono scartati tutti i *tweet* contenenti meno di tre termini –parole e *hashtag*– poiché si assume che contengano una quantità di informazioni trascurabile.

6.2 Analisi Esplorative

La collezione filtrata è formata da $D = 8,895$ *tweet* pubblicati dai $U = 101$ utenti considerati attivi nell'intervallo di tempo considerato. In media a ogni utente sono associati 88.0693 *tweet* e in Figura 6.1 si osserva che la maggior parte degli utenti ha pubblicato tra i 32 –valore minimo[8]– e i 130 *tweet*, mentre una minoranza risulta essere molto più attiva nella piattaforma con una produzione compresa tra i 200 e i 600 *tweet*.

[8]La collezione può contenere meno di 70 *tweet* di un utente poiché, dopo aver selezionato i 101 utenti attivi, è stato effettuato un ulteriore filtro che ha eliminato parte dei documenti.

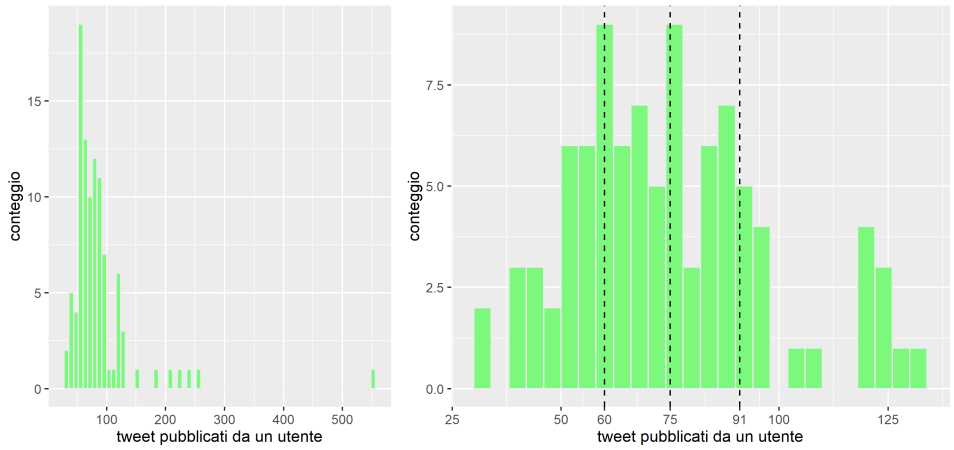


Figura 6.1: Distribuzione del numero di *tweet* pubblicati per ogni utente; le linee verticali sono i quartili della distribuzione.

Per quanto riguarda il contenuto dei testi, la collezione contiene $N = 79,721$ parole e $L = 715$ *hashtag*, la cui suddivisione all'interno dei *tweet* è rappresentata nelle due distribuzioni riportate in Figura 6.2. Partendo dagli *hashtag*, si osserva che sono assenti in quasi la totalità dei *tweet* (94.83%) e che si hanno al massimo cinque *hashtag* all'interno di uno stesso *tweet*. Proseguendo, in media un *tweet* contiene 8.9626 parole e la maggior parte di essi ne contiene tra le 3 e le 25; in particolare, per una migliore interpretabilità, in figura si mostrano solo i conteggi dei *tweet* contenenti 30 o meno parole, evidenziando un elevato numero di *tweet* corti, contenenti meno di 10 parole (63.94%). Le linee verticali tratteggiate rappresentano i quartili delle distribuzioni, mentre quella non tratteggiata è la media; i quartili non sono riportati nella distribuzione del numero di *hashtag* poiché sono tutte e tre pari a zero.

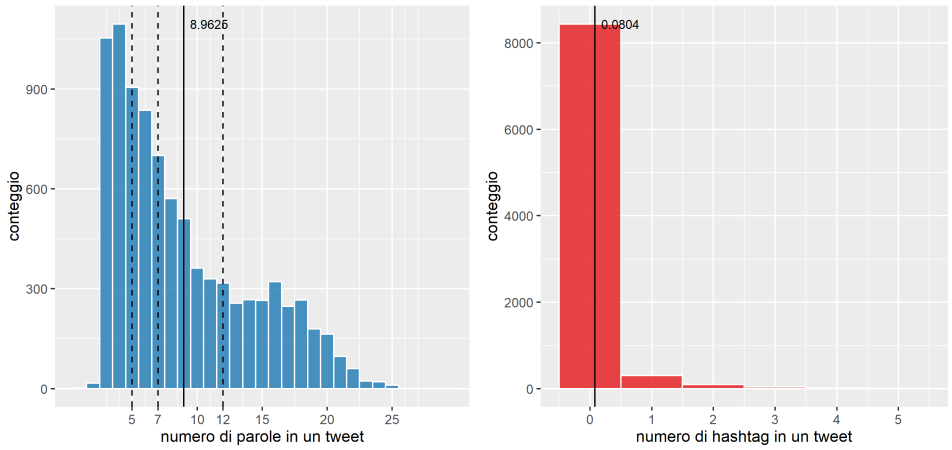


Figura 6.2: Istogramma del numero di parole (sinistra) e del numero di *hashtag* (destra) contenuti in ogni *tweet*.

L'ampiezza del vocabolario delle parole è $V = 24,115$ e l'ampiezza del vocabolario degli *hashtag* è $H = 123$: uno strumento utile per visualizzare la frequenza con cui gli elementi di un vocabolario occorrono in un testo o una collezione di testi è il *wordcloud*. In Figura 6.3a si riporta il *wordcloud* delle parole che occorrono almeno 100 volte nella collezione di *tweet*, mentre in Figura 6.3b si riportano tutti gli *hashtag*; in entrambe le immagini i colori sono casuali, mentre la grandezza di un termine è proporzionale al suo numero di occorrenze (*term frequency*). Come è ragionevole aspettarsi in una collezione costruita a partire dalla tematica *covid*, sia le parole sia gli *hashtag* ad alta frequenza sono legati ad esso.

(a) *Wordcloud* delle parole. **(b)** *Wordcloud* degli *hashtag*.

Figura 6.3: *Wordcloud* delle parole e degli *hashtag* contenuti nella collezione di *tweet*.

Si introduce ora il concetto di distribuzione della collezione (*corpus distribution*) per rappresentare la frequenza con cui compaiono gli elementi in un vocabolario: avendo due vocabolari, si definiscono due distribuzioni della collezione. Questa è la stessa rappresentazione dei topic nei *topic model*; in particolare, è come se si stesse considerando un unico topic per l'intera collezione (Boyd-Graber et al., 2014).

La distribuzione sulle parole della collezione, $\boldsymbol{\phi}^{\mathcal{C}}$, è un vettore $V \times 1$ il cui v-mo elemento è dato da

$$\phi_v^{\mathcal{C}} = \frac{\beta_v^V + \sum_{u=1}^{U} \sum_{d=1}^{D_u} \sum_{n=1}^{N_{ud}} \mathbb{1}_{w_{udn}=v}}{\sum_{v=1}^{V} \beta_v^V + N};$$

dove N è il numero di parole nella collezione e $\boldsymbol{\beta}^V$ è lo stesso parametro utilizzato per stimare i *topic model*. Analogamente, la distribuzione sugli *hashtag* della collezione, $\boldsymbol{\psi}^{\mathcal{C}}$, è un vettore $H \times 1$ il cui v-mo elemento è dato da

$$\psi_h^{\mathcal{C}} = \frac{\beta_h^H + \sum_{u=1}^{U} \sum_{d=1}^{D_u} \sum_{l=1}^{L_{ud}} \mathbb{1}_{h_{udl}=h}}{\sum_{h=1}^{H} \beta_h^H + L};$$

dove L è il numero di *hashtag* nella collezione e $\boldsymbol{\beta}^H$ è lo stesso parametro utilizzato per stimare i *topic model*. Le parole e gli *hashtag* a cui sono associate le probabilità più alte in $\boldsymbol{\phi}^{\mathcal{C}}$ e $\boldsymbol{\psi}^{\mathcal{C}}$, ovvero le *top word* e i *top hashtag* delle due distribuzioni, sono riportati in Figura 6.4.

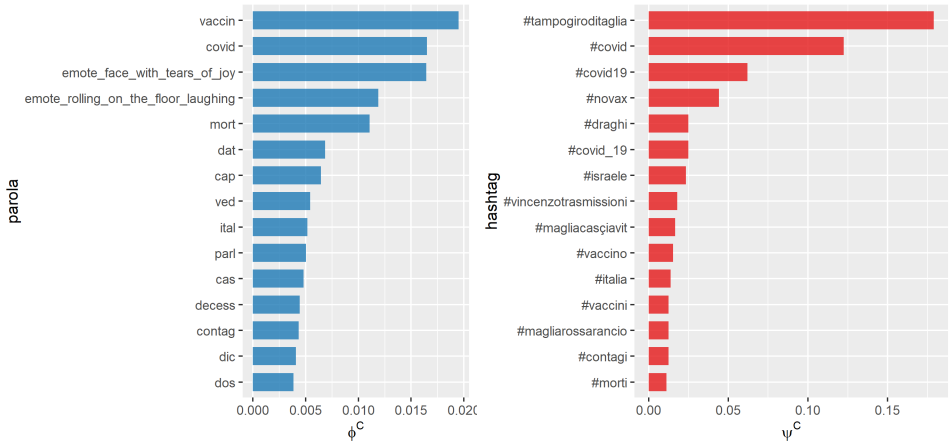


Figura 6.4: Lista delle 15 *top word* della distribuzione sulle parole, $\boldsymbol{\phi}^{\mathcal{C}}$, e lista dei 15 *top hashtag* della distribuzione sugli *hashtag*, $\boldsymbol{\psi}^{\mathcal{C}}$, della collezione di *tweet*.

Si noti infine che le *top word* emote_face_with_tears_of_joy e emote_rolling_on_the_floor_laughing non sono presenti nella *wordcloud* delle parole esclusivamente a causa della loro eccessiva lunghezza.

6.3 Analisi Quantitative

Il modello proposto è valutato sulla collezione filtrata di *tweet* e confrontato sia quantitativamente sia qualitativamente con i tre suoi casi particolari introdotti nel Capitolo 3: *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA*. L'inferenza approssimata dei quattro *topic model* è effettuata attraverso dei *Collapsed Gibbs Sampler* con lo stesso numero di iterazioni, in modo da poter confrontare i risultati ottenuti a partire da catene di Markov della stessa lunghezza. Emulando quanto fatto in F. Zhao et al., 2016, si fissa il numero di iterazioni degli algoritmi a 300: le prime 199 iterazioni sono scartate poiché considerate di *burn-in*, mentre delle rimanenti si considera una iterazione ogni dieci, ottenendo quindi un campione di 11 osservazioni su cui applicare le tecniche *Monte Carlo*.

6.3.1 Parametri

Come già accennato nella sottosezione 2.3.2, il parametro più importante in un qualsiasi *topic model parametrico* è il numero di topic T poiché influenza fortemente sia la qualità dei topic sia l'approccio da seguire per l'interpretazione dei risultati; inoltre, T influenza in maniera non indifferente il tempo necessario per effettuare l'inferenza. Utilizzando l'implementazione efficiente della *Latent Dirichlet Allocation* contenuta nella libreria `topicmodels` (Grün & Hornik, 2011) di *R*, si è stimata la *Latent Dirichlet Allocation* con diversi T, ottenendo i risultati riportati in Tabella 6.1. Infine, si è scelto di stimare i quattro *topic model* con 10, 30, 70, 100, considerando $T = 30$ come valore di riferimento –dal momento che si ha TC-PMI massima in $T = 30$– e i rimanenti come alternative ragionevoli per rendere possibile una valutazione dei modelli al variare del numero di topic.

topic	10	20	30	40
TC-PMI	-238.3981	-235.7742	**-227.8729**	-229.1626

topic	50	60	70	80
TC-PMI	-233.4593	-236.3656	-237.4954	-242.9405

topic	90	100	110	120
TC-PMI	-245.1213	-248.9779	-251.2230	-249.0772

Tabella 6.1: Risultati delle analisi preliminari per la selezione del numero dei topic T; TC-PMI è calcolato su $N = 10$ *top word* di ogni topic.

I rimanenti parametri sono selezionati in modo tale da avere delle distribuzioni a priori non informative: tutti i parametri delle distribuzioni Beta sono fissati a 1,[9] mentre per le distribuzioni sui topic, sulle parole e sugli *hashtag* si considerano distribuzioni di Dirichlet simmetriche. Per fissare i parametri di queste ultime si seguono le indicazioni di F. Zhao et al., 2016, che per le distribuzioni sui topic riprende a sua volta i valori consigliati da Griffiths e Steyvers, 2004 e Steyvers e Griffiths, 2007:

$$\begin{aligned}\boldsymbol{\alpha}^* &= \frac{50}{T}\mathbf{1}_T \\ \boldsymbol{\alpha} &= \frac{50}{T}\mathbf{1}_T \\ \boldsymbol{\beta}^V &= \frac{1}{10}\mathbf{1}_V \\ \boldsymbol{\beta}^H &= \frac{1}{10}\mathbf{1}_H\end{aligned}$$

Infine, il *weak topic smoothing prior* α_0 è fissato a 10^{-7}, come suggerito in Lin et al., 2014.

6.3.2 Convergenza

A causa dell'elevato costo computazionale e dell'elevata dimensionalità della struttura latente dei *topic model*, spesso l'utilizzo di metodi *MCMC* non è accompagnato da un'approfondita analisi della convergenza: questa tende

[9] Si noti che una distribuzione Beta di parametri $(1, 1)$, $Beta(1, 1)$, coincide con una distribuzione continua uniforme sull'intervallo unitario, $U(0, 1)$.

ad essere messa in secondo piano dal tempo limitato a disposizione per far eseguire gli algoritmi (Cohen, 2019).

Essendo in un contesto ad alta dimensionalità in cui non è possibile applicare le diagnostiche standard di convergenza, si utilizza una funzione scalare delle variabili del *topic model* per valutare graficamente se la catena ha superato la fase di *burn-in*. In questa tesi, si considera il logaritmo della distribuzione congiunta di variabili osservate e latenti poiché è semplice e rapido da calcolare in ogni stato della catena; in Appendice B sono riportate le formulazioni utilizzate per tracciare i grafici nelle Figure 6.5, 6.6, 6.7 e 6.8. Le linee tratteggiate sono i logaritmi delle distribuzioni congiunte calcolate nei vari stati delle catene, mentre le linee non tratteggiate sono le *stime Monte Carlo* della stessa quantità calcolate al variare del numero di iterazioni.

Formalmente, sia $\ell(\mathbf{U^{(i)}})$ il logaritmo della distribuzione congiunta di variabili osservate e latenti calcolato nell'i-mo stato di una catena, $\mathbf{U^{(i)}}$; la stima *Monte Carlo* di $\ell(\cdot)$ calcolata sui primi i stati della catena è dato da

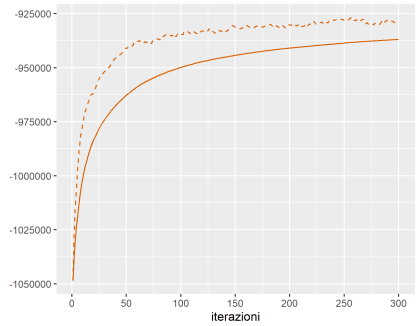
$$\frac{1}{i}\sum_{j=1}^{i}\ell(\mathbf{U^{(j)}})$$

Per la *legge dei grandi numeri*, questa quantità dovrebbe stabilizzarsi e raggiungere un *plateau* per i abbastanza elevato: il problema della convergenza è identificare questo i.

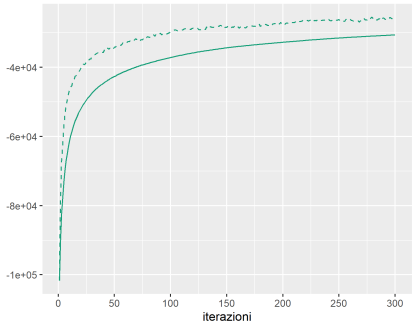
Tenendo conto del numero molto ridotto di iterazioni, scelta necessaria a causa di un'implementazione non efficiente, si ottengono dei risultati abbastanza soddisfacenti dal momento che si osserva una stabilizzazione dei logaritmi delle distribuzioni congiunte; tuttavia le stime *Monte Carlo*, non avendo ancora raggiunto un *plateau*, suggeriscono che le catene debbano ancora raggiungere la convergenza.

6.3.3 Metriche di Valutazione

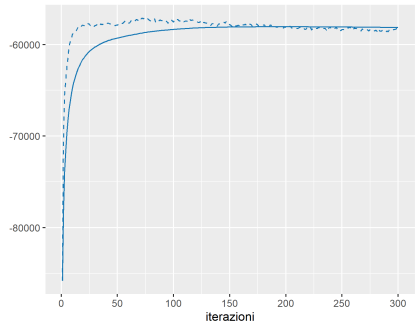
La natura non supervisionata dei *topic model* fornisce loro grande flessibilità, ma allo stesso tempo l'assenza di annotazioni rende più difficile la loro valutazione. Agli albori del *topic modeling* la bontà di adattamento viene valutata con la *perplexity* o la *held-out likelihood*; tuttavia presto vengono abbandonate in favore delle metriche di *topic coherence* poiché, oltre a non fornire



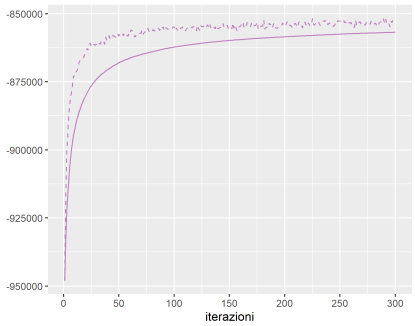
LDA con 10 topic.



Twitter-LDA con 10 topic.



Hashtag-LDA con 10 topic.

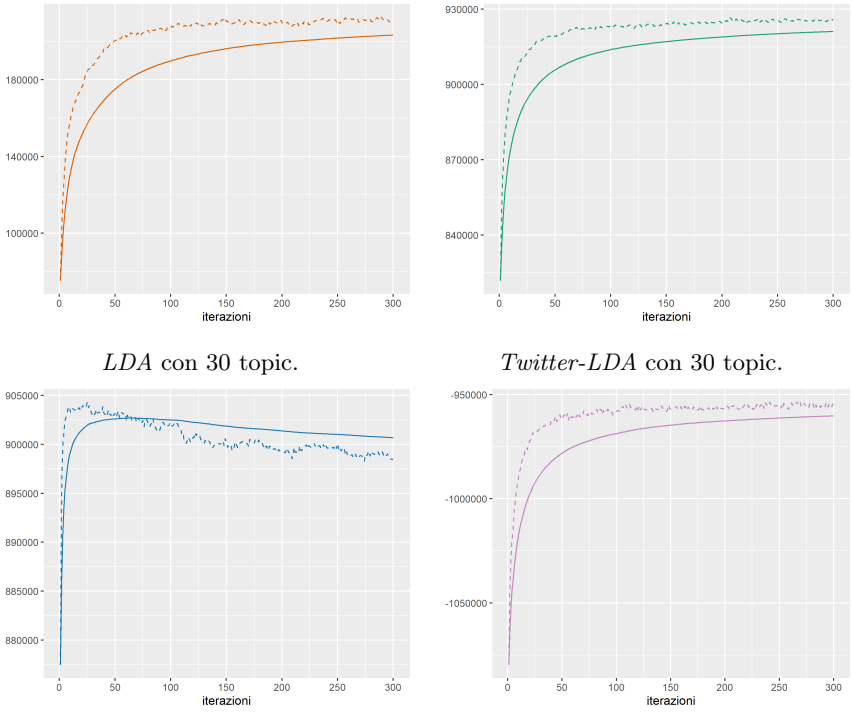


Modello proposto con 10 topic.

Figura 6.5: *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i.

una valutazione quantitativa della struttura latente, i.e. non posso fornire informazioni sulla qualità dei topic identificati, Chang et al., 2009 dimostrano che le due quantità sono correlate negativamente con l'interpretabilità del modello. Più nello specifico, le metriche di *topic coherence*, inizialmente basate sul giudizio umano (Chang et al., 2009), diventano l'approccio più comune per valutare *topic model* con l'introduzione di metodi automatici (Newman et al., 2010, Mimno et al., 2011). Le metriche di *topic coherence* si basano sull'intuizione secondo cui un umano percepisce un topic informativo e coerente se le sue *top word* tendono ad essere utilizzate congiuntamente dal momento che sono in qualche modo legate tra loro (Boyd-Graber et al., 2014).

Boyd-Graber et al., 2014 propongono una formulazione generale per calcolare la *topic coherence* di un singolo topic; siano $\mathbf{w}^{(t)} = (w_1^{(t)}, \ldots, w_N^{(t)})$ le N parole a cui sono associate le probabilità $\phi_{t,v}$ più alte per t fissato e



LDA con 30 topic. *Twitter-LDA* con 30 topic.

Hashtag-LDA con 30 topic. Modello proposto con 30 topic.

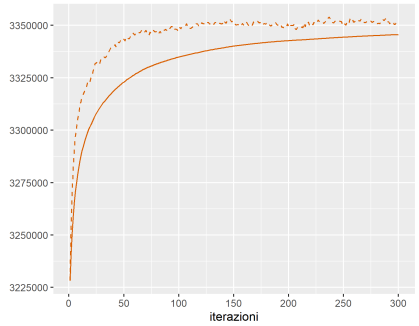
Figura 6.6: *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i.

$v = 1, \ldots, V$, e sia $f(\cdot)$ una funzione che misura l'associazione tra due parole, allora la *topic coherence* del topic t è definita come:

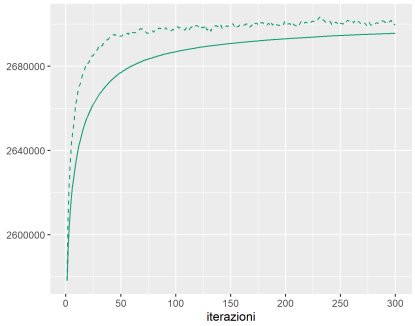
$$\text{TC-}f(\mathbf{w}^{(t)}) = \frac{2}{N^2 - N} \sum_{i=2}^{N} \sum_{j=1}^{i-1} f(w_i^{(t)}, w_j^{(t)})$$

Considerando diverse funzioni $f(\cdot)$ è possibile ottenere diverse metriche di *topic coherence* a partire dalla formulazione generale sopra esposta; in questa tesi si considerano le tre metriche riportate in Boyd-Graber et al., 2014:

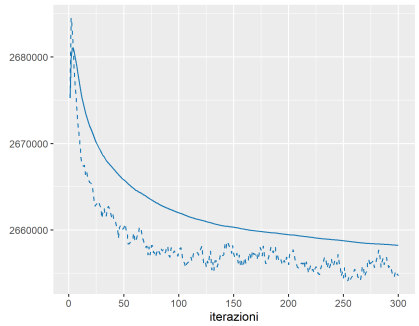
- TC-PMI (Newman et al., 2010)
  In questo caso la funzione $f(\cdot)$ è la *informazione mutua puntuale* (*Pointwise Mutual Information*, *PMI*), una misura di associazione che quantifica la discrepanza tra la probabilità di co-occorrenza delle due parole data la loro distribuzione congiunta rispetto alla stessa quantità sotto



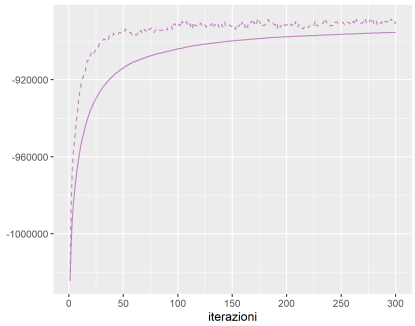
LDA con 70 topic.



Twitter-LDA con 70 topic.



Hashtag-LDA con 70 topic.



Modello proposto con 70 topic.

Figura 6.7: *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i.

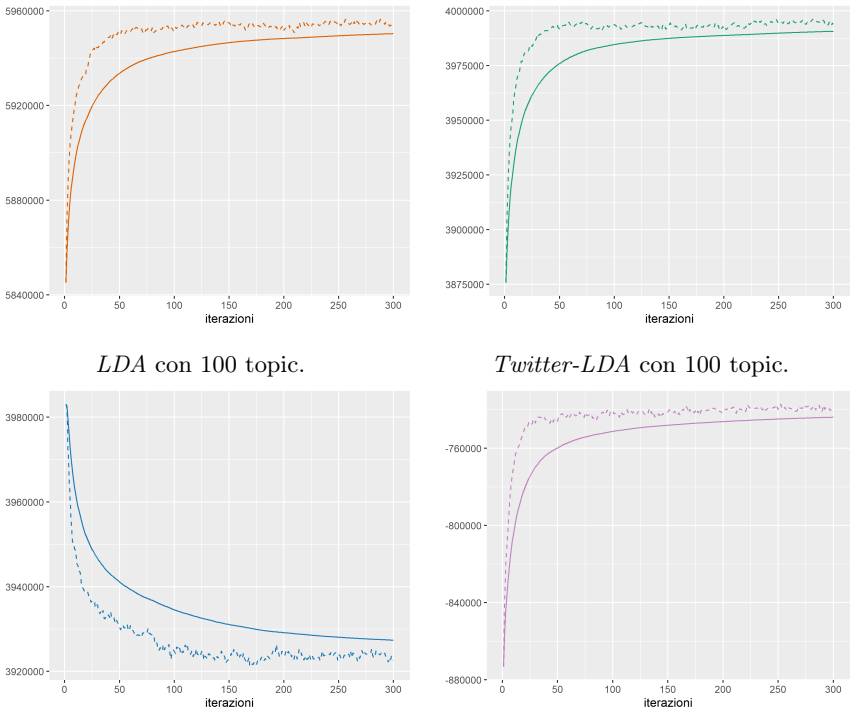
l'assunzione di indipendenza. In zero si ha indipendenza tra le parole, valori positivi indicano che le parole tendono a co-occorrere e valori negativi che le due parole tendono a non comparire negli stessi documenti. La metrica di *topic coherence* TC-PMI è così definita:

$$\text{TC-PMI}(\mathbf{w}^{(t)}) = \frac{2}{N^2 - N} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \text{PMI}(w_i^{(t)}, w_j^{(t)})$$

dove

$$\text{PMI}(w_i^{(t)}, w_j^{(t)}) = \begin{cases} log\left(\dfrac{p(w_i^{(t)}, w_j^{(t)})}{p(w_i^{(t)})p(w_j^{(t)})}\right) & \text{se } p(w_i^{(t)}, w_j^{(t)}) > 0 \\ 0 & \text{altrimenti} \end{cases}$$

$p(w_i, w_j)$ è il numero di documenti in cui la parola w_i e la parola w_j co-occorrono sul totale dei D documenti considerati e $p(w_i)$ è il numero



LDA con 100 topic. *Twitter-LDA* con 100 topic.

Hashtag-LDA con 100 topic. Modello proposto con 100 topic.

Figura 6.8: *Trace plot* del logaritmo della distribuzione congiunta di variabili osservate e latenti (tratteggiato) e stima *Monte Carlo* della stessa quantità calcolata sulle prime i iterazioni, al variare del numero di iterazioni i.

di documenti in cui compare la parola w_i sul totale dei D documenti considerati.

- TC-LCP (Mimno et al., 2011)
  In questo caso la funzione $f(\cdot)$ è la *Log Conditional Probability* (LCP), una misura di associazione molto simile a PMI nata dall'intuizione che ciò che conta non è la differenza tra la distribuzione congiunta e le marginali, ma la probabilità condizionale di osservare ogni parola data un'altra con peso maggiore all'interno del topic. In zero si ha coerenza massima, quindi come per TC-PMI è preferibile osservare valori alti di TC-LCP. La metrica di *topic coherence* TC-LCP è così definita:

$$\text{TC-LCP}(\mathbf{w}^{(t)}) = \frac{2}{N^2 - N} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \text{LCP}(w_i^{(t)}, w_j^{(t)})$$

$$= \frac{2}{N^2 - N} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \left(\frac{N(w_i^{(t)}, w_j^{(t)}) + \beta}{N(w_j^{(t)})} \right)$$

dove $N(w_i^{(t)}, w_j^{(t)})$ è il numero di documenti in cui la parola $w_i^{(t)}$ e la parola $w_j^{(t)}$ co-occorrono e $N(w_i^{(t)})$ è il numero di documenti in cui compare la parola $w_i^{(t)}$; $\beta > 0$ è uno *smoothing count*, solitamente pari a 1, necessario per evitare di avere l'argomento del logaritmo uguale a zero.

- TC-NZ (Boyd-Graber et al., 2014)
  In questo caso la funzione $f(\cdot)$ si limita a contare quante volte due parole non co-occorrono mai nella collezione di riferimento; TC-NZ è una percentuale che indica quante tra le *top word* di un topic non co-occorrono mai. A differenza delle due metriche precedenti, è preferibile osservare TC-NZ bassi. La metrica di *topic coherence* TC-NZ è così definita:

$$\text{TC-NZ}(\mathbf{w}^{(t)}) = \frac{2}{N^2 - N} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \text{NZ}(w_i^{(t)}, w_j^{(t)})$$

$$= \frac{2}{N^2 - N} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \mathbb{1}_{N(w_i^{(t)}, w_j^{(t)})=0}$$

  dove $N(w_i^{(t)}, w_j^{(t)})$ è il numero di documenti in cui la parola $w_i^{(t)}$ e la parola $w_j^{(t)}$ co-occorrono.

Le quantità $p()$ e $N()$ possono essere calcolate sulla stessa collezione su cui è stato stimato il modello o su una collezione esterna, e le co-occorrenze su interi documenti o su una finestra di parole. In questa tesi, per semplicità, si utilizza la stessa collezione e, data la brevità dei documenti, ovvero i *tweet*, si considerano le co-occorrenze su interi documenti. Inoltre, si considerano le $N = 10$ parole a cui sono associate le probabilità $\phi_{t,v}$ più alte.

A questo punto, per ottenere un unico valore che fornisce un'indicazione della qualità complessiva dei topic identificati da un modello, si effettua semplicemente la media delle *topic coherence* dei suoi topic:

$$\frac{1}{T} \sum_{t=1}^{T} \text{TC-}f(\mathbf{w}^{(t)})$$

Nella prime tre colonne della Tabella 6.2 si riportano le tre metriche di *topic coherence* calcolate sui $\boldsymbol{\phi}_{1:T}$ dei 16 *topic model* stimati. L'ultima colonna

contiene la distanza media delle distribuzioni sulle parole dei topic, $\phi_{1:T}$, dalla distribuzione sulle parole della collezione, $\phi^{\mathcal{C}}$. La distanza tra distribuzioni è misurata con la *divergenza di Jensen-Shannon* e può essere usata come indicatore della qualità di un topic: se la distribuzione sulle parole di un topic è vicina alla distribuzione della collezione, allora il topic è percepito come inutile o troppo generale (AlSumait et al., 2009). Le quantità riportate in Tabella 6.2 sono rappresentate in Figura 6.9.

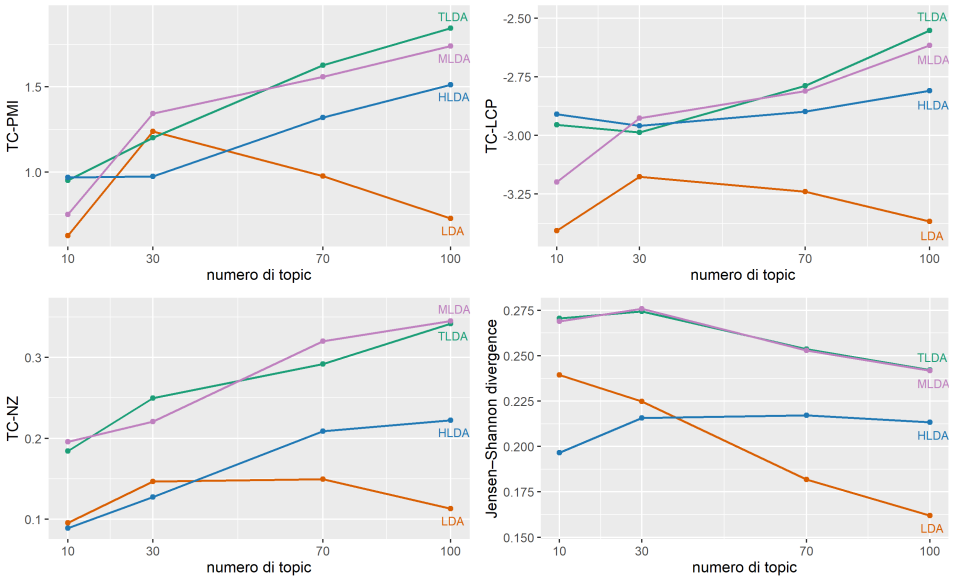


Figura 6.9: Rappresentazione grafica di TC-PMI, TC-LCP, TC-NZ con $N = 10$ dei 16 *topic model* stimati; il grafico in basso a destra mostra la distanza media tra le righe di $\phi_{1:T}$ e $\phi^{\mathcal{C}}$, misurata con la *divergenza di Jensen-Shannon*. I *topic model* considerati sono: *Latent Dirichlet Allocation* (LDA), *Twitter-LDA* (TLDA), *Hashtag-LDA* (HLDA) e il modello proposto (MLDA).

Come è ragionevole aspettarsi, complessivamente la *Latent Dirichlet Allocation* risulta essere il modello nettamente meno adatto per la collezione di *tweet*; in particolare, a parità di topic ha quasi sempre il TC-PMI e TC-LCP più basso, e le distribuzioni sulle parole dei suoi topic sono in media le più vicine a quella della collezione. Per gli altri tre modelli si osserva che, all'aumentare del numero di topic, aumentano TC-PMI e TC-LCP, tuttavia diminuisce la distanza media delle distribuzioni sulle parole dei topic da quella della collezione: si ha quindi una coerenza maggiore non perché i topic riescono a identificare con più efficacia le parole caratterizzanti delle

tematiche della collezione, ma perché tutte le distribuzioni sulle parole dei topic tendono ad avvicinarsi alla distribuzione della collezione, avendo tra le loro *top word* le parole a frequenza più alta. In questa particolare situazione la metrica di *topic coherence* complessiva di un *topic model* risulta alta, tuttavia i suoi topic risultano essere vaghi e non adatti per isolare le varie tematiche della collezione, come discusso in Boyd-Graber et al., 2014. In generale, è necessario valutare un trade-off tra le metriche di *topic coherence* e la distanza media: favorendo eccessivamente uno dei due indicatori, si rischia di stimare una struttura latente non informativa.

Infine, è interessante notare che, anche al variare del numero di topic, *Twitter-LDA* e il modello proposto, ovvero i due *topic model* che considerano parole di sottofondo nel processo generativo, presentano indicatori quasi coincidenti e migliori rispetto ai due modelli rimanenti: questa è un'indicazione del fatto che considerare una distinzione tra parole generate a partire da un topic e parole di sottofondo in questa collezione permette di ottenere topic più coerenti e informativi, le cui distribuzioni sulle parole tendono ad essere più lontane dalla distribuzione della collezione.

6.4 Analisi Qualitative

In questa sezione si considera il modello proposto con $T = 30$ e si mostra come sfruttare le stime a posteriori per analizzare le caratteristiche della collezione, dei documenti, degli utenti e dei 30 topic latenti identificati dal modello. Si noti che l'obiettivo di questa sezione non è analizzare la collezione di *tweet* sui cui è stato applicato il modello, ma mostrare come è possibile interpretare i parametri contenuti in **par** e le variabili latenti **x**.

6.4.1 Tipo dei Documenti

Per ogni documento ud si ha a disposizione una probabilità x_{ud} calcolata come

$$x_{ud} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} x_{ud}^{(i)}$$

dove $\mathcal{M} = \{200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300\}$ sono gli indici degli stati della catena di Markov considerati per calcolare le stime *Monte Carlo* a posteriori. Si assume che un documento ud tratti di più topic

	TC-PMI	TC-LCP	TC-NZ	JS
LDA (10)	0.6259	-3.4068	0.0956	0.2393
LDA (30)	1.2389	-3.1759	0.1467	0.2249
LDA (70)	0.9757	-3.2396	0.1495	0.1817
LDA (100)	0.7269	-3.3673	0.1129	0.1620
Twitter-LDA (10)	0.9498	-2.9537	0.1844	0.2705
Twitter-LDA (30)	1.2024	-2.9867	0.2496	0.2743
Twitter-LDA (70)	1.6274	-2.7870	0.2921	0.2536
Twitter-LDA (100)	**1.8443**	**-2.5519**	0.3418	0.2422
Hashtag-LDA (10)	0.9676	-2.9089	**0.0889**	0.1966
Hashtag-LDA (30)	0.9734	-2.9582	0.1274	0.2156
Hashtag-LDA (70)	1.3200	-2.8978	0.2089	0.2172
Hashtag-LDA (100)	1.5133	-2.8095	0.2222	0.2133
Modello proposto (10)	0.7502	-3.1981	0.1956	0.2689
Modello proposto (30)	1.3433	-2.9257	0.2207	**0.2758**
Modello proposto (70)	1.5590	-2.8099	0.3200	0.2529
Modello proposto (100)	1.7395	-2.6162	0.3451	0.2417

Tabella 6.2: TC-PMI, TC-LCP e TC-NZ con $N = 10$ dei 16 *topic model* stimati; l'ultima colonna contiene la distanza media tra le distribuzioni sulle parole dei topic e la distribuzione sulle parole della collezione, misurata con la *divergenza di Jensen-Shannon*.

se $x_{ud} > 0.5$; nella seguente tabella si osserva che solo quattro documenti all'interno della collezione di *tweet* trattano di più topic:

x_{ud}	0	0.0909	0.2727	0.63636	1
frequenza	8885	5	1	1	3

Coerentemente a ciò, si ha che tutte le probabilità contenute in $\boldsymbol{\pi}_{1:T}$ assumono valori estremamente bassi, indicando appunto che tendenzialmente tutti gli utenti pubblicano *tweet* non eccessivamente elaborati che trattano di un unico topic.

Si noti che anche i *tweet* più lunghi all'interno della collezione non utilizzano un'ampia varietà di termini al loro interno e presentano pochi termini –*emoji* in particolare– ripetuti molte volte. In generale, termini uguali all'interno di uno stesso documento tendono ad essere generati dallo stesso topic, quindi in questo caso la procedura d'inferenza tende ad associare $x_{ud} = 0$

anche a documenti lunghi con poche parole distinte ripetute molte volte.

6.4.2 Rappresentazione dei Documenti come Mistura di Topic

I quattro documenti che trattano di più topic possono essere interpretati come mistura di topic; è quindi possibile adottare lo stesso approccio utilizzato nella *LDA*. Un metodo molto semplice, e allo stesso tempo efficace, per rappresentare un singolo documento ud come mistura è un grafico a barre in cui l'altezza della t-ma barra è proporzionale al peso del topic t all'interno del documento ud, $\theta_{ud,t}$. In Figura 6.10 si osserva che nei primi tre documenti pochi topic hanno un peso elevato, mentre nel quarto, rappresentato in basso a destra, molti topic superano la soglia $\frac{1}{30}$, ovvero il valore che si assume a priori per i pesi dei topic all'interno di un documento.

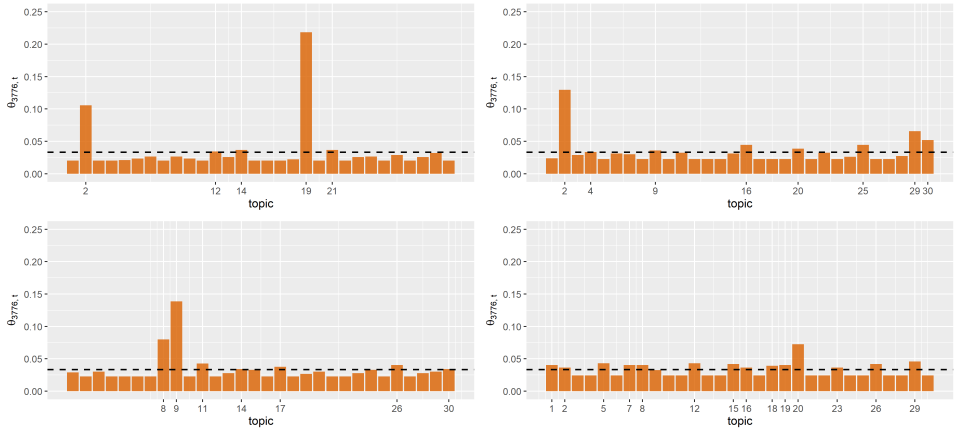


Figura 6.10: Rappresentazione dei quattro documenti delle collezione che trattano di più topic come misture di topic; la linea orizzontale ha come ordinata $\frac{1}{30}$.

6.4.3 Preferenze degli Utenti

La distribuzione sui topic, $\boldsymbol{\theta}^*_u$, di un utente u ha un'interpretazione analoga a quella adottata per la distribuzione sui topic, $\boldsymbol{\theta}_{ud}$ di un documento ud; si omette quindi questa parte per presentare un modo alternativo per valutare le distribuzioni sui topic degli utenti nel loro complesso. L'approccio che verrà descritto di seguito può essere applicato anche ai documenti; infatti, in generale, le distribuzioni sui topic e sugli utenti possono essere trattate

allo stesso modo, tenendo però presente che la distribuzione sui topic di un documento ha significato solo se il documento tratta di più topic.

Un primo modo per valutare un topic è contare quante volte esso è il preferito di un utente, ovvero assume il valore massimo all'interno della distribuzione sui topic dell'utente; formalmente, l'indice k del topic preferito dell'utente u è dato da

$$k = \underset{t \in \{1,\dots,T\}}{\operatorname{argmax}} \{\theta^*_{u,t}\}$$

Alternativamente, è possibile contare per quanti utenti ogni topic risulta essere importante; seguendo la definizione di Boyd-Graber et al., 2014, un topic t può essere considerato importante per un utente u se il suo peso supera una soglia fissata ε, ovvero se $\theta^*_{u,t} > \varepsilon$. Intuitivamente, fissato $\varepsilon = 0.2$, si ha che un topic t è importante per un utente u se almeno il 20% –circa– dei *tweet* dell'utente u ha il topic t come topic principale.

I due approcci sono rappresentati nella Figura 6.11 e nella Figura 6.12: osservando il primo grafico, il topic 7 risulta essere il più importante (per 10.89% degli utenti), seguito dal 29 (8.91%) e poi i topic 23, 24 e 30 a pari merito (6.93%); osservando il secondo grafico emerge che, anche se un topic è il preferito di un utente, non è necessariamente trattato in almeno il 20% dei suoi *tweet*.

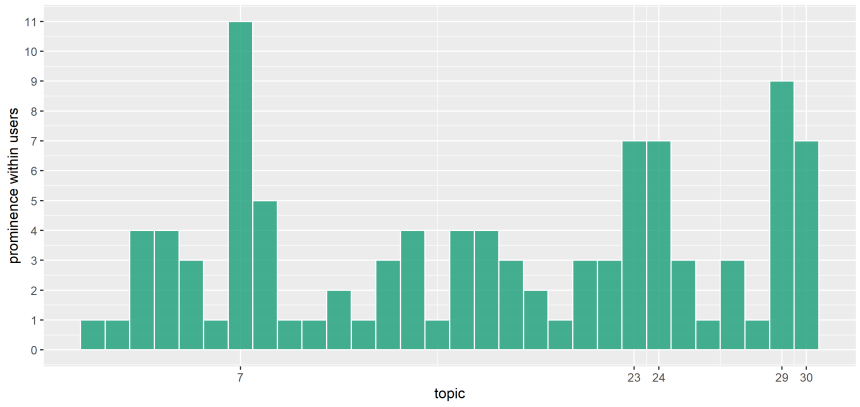


Figura 6.11: Distribuzione del numero di utenti per cui il topic considerato è quello preferito per ogni topic; l'indice k del topic preferito dell'utente u è dato da $k = \text{argmax}_{t \in \{1,\dots,T\}}\{\theta^*_{u,t}\}$.

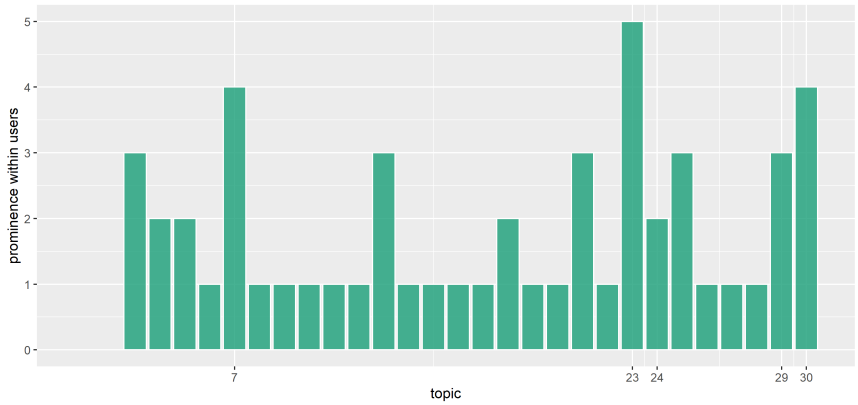


Figura 6.12: Distribuzione del numero di utenti per cui il topic considerato è importante per ogni topic; un topic t è importante per l'utente u se $\theta_{u,t} \geq 0.2$.

6.4.4 Doppia Rappresentazione dei Topic

Si procede quindi all'interpretazione dei topic sfruttando la loro doppia rappresentazione, sia come distribuzione sulle parole, $\boldsymbol{\phi}_{1:T}$, sia come distribuzione sugli *hashtag*, $\boldsymbol{\psi}_{1:T}$. Solitamente, nei *topic model* in cui non è presente una distinzione tra parole e *hashtag* –ad esempio, *LDA* e *Twitter-LDA*– la tematica di ogni topic t è ricavata esclusivamente dalle sue *top word*, intese come gli elementi v del vocabolario delle parole a cui corrispondono i $\phi_{t,v}$ più alti. Nel *topic model* proposto in questa tesi, avendo a disposizione anche una distribuzione sugli *hashtag*, è possibile determinare le tematica a partire sia dalle *top word* sia dai *top hashtag*, intesi come gli elementi h del vocabolario degli *hashtag* a cui corrispondono i $\psi_{t,h}$ più alti.

Nella sezione precedente sono stati identificati i cinque topic che più frequentemente risultano essere i preferiti dagli utenti; un approccio ragionevole per l'esplorazione delle tematiche della collezione è partire proprio da questi topic dal momento che rappresentano i topic preferiti di più del 40% degli utenti considerati — 40.59% per essere precisi. Si noti che in questo caso, dato il numero relativamente ridotto di topic, sarebbe possibile analizzare tutti i topic uno a uno. Tuttavia, un'analisi di tutti i topic sarebbe di difficile attuazione nel caso in cui si considerasse un numero estremamente elevato di topic, ad esempio nell'ordine delle centinaia o, in casi estremamente rari, nell'ordine delle migliaia.

Topic 7 La tematica di questo topic sembra essere legata alle misure adottate dal Governo per contenere il covid; in particolare, è interessante notare che le *top word* permettono di identificare la tematica (pass, gp, gren, vaccin), mentre i *top hashtag* evidenziano il tono, tutt'altro che amichevole, del topic (#salvinidimettiti, #draghistan, #draghiingalera, #draghivattene). Le *top word* e i *top hashtag* sono rappresentati in Figura 6.13.

Topic 23 I *top hashtag* #novax, #vaccinatevi_e_basta, #untori, #nobrain suggeriscono che questo topic sia tipico dei *tweet* in cui si accusano i no vax di basare la loro scelta di non vaccinarsi su notizie la cui fonte non è attendibile (#fakenews, #miocuginonews). Inoltre, è interessante notare che la "doppia" rappresentazione è utile per identificare la tematica, identificazione che non sarebbe possibile considerando solo le *top word*. Le *top word* e i *top hashtag* sono rappresentati in Figura 6.14.

Topic 24 La tematica di questo topic sembra essere legata all'obbligo vaccinale degli over 50 e alle sanzioni economiche legate ad esso (mult, pag, sol, controll); i *top hashtag* sono coerenti, infatti si osservano tre variazioni delle stesso concetto con #vaccini, #vaccino e #vaccinati. Le *top word* e i *top hashtag* sono rappresentati in Figura 6.15.

Topic 29 La tematica di questo topic è la situazione della pandemia nel mondo, in particolare le *top word* osservate sono proprie dei *tweet* informativi il cui scopo è aggiornare gli utenti della piattaforma *Twitter* sulla situazione dei contagi (contag, numer), dei morti (mort, decess) e della diffusione del virus e delle sue varianti (omicron, variant, virus). I *top hashtag* sono coerenti, infatti si osservano ad esempio #draghi, #europa, #israele. Le *top word* e i *top hashtag* sono rappresentati in Figura 6.16.

Topic 30 Questo topic è molto simile a quello precedente, ma sembra essere più incentrato sugli aggiornamenti di routine della situazione dei contagi in Italia; in particolare, anche qui si osservano parole tipiche dei *tweet* informativi il cui scopo è aggiornare gli utenti della piattaforma Twitter sulla situazione dei contagi: mort, decess, ospedal, posit, numer, ricover, tampon. I primi tre *top hashtag* sono coerenti con questa interpretazione dal momento

che solitamente i *tweet* che espongono la situazione dei contagi contengono sempre gli *hashtag* #covid19, #covid_19, #covid o simili. Le *top word* e i *top hashtag* sono rappresentati in Figura 6.17.

Topic 22 Infine, un topic interessante è il 22 poiché le sue prime sei *top word* sono delle *emoji*, è quindi ragionevole affermare che i *tweet* caratterizzati da questo topic contengono principalmente *emoji*. Le *top word* e i *top hashtag* del topic 22 è rappresentato in Figura 6.18.

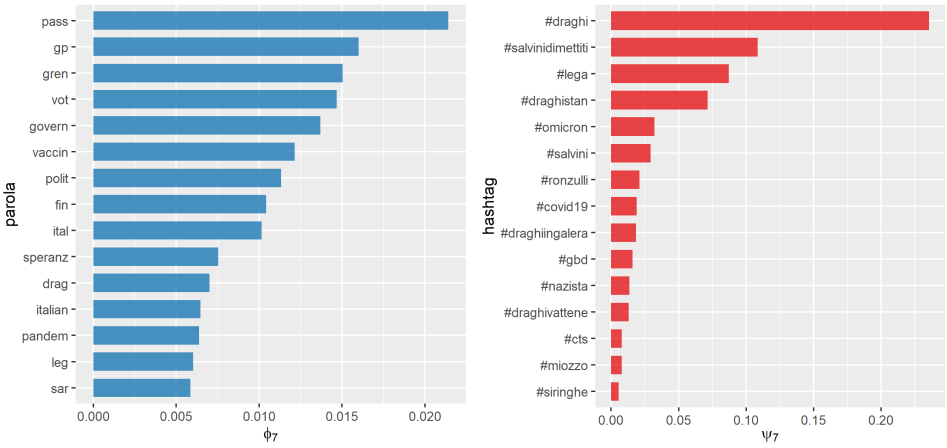


Figura 6.13: Lista delle 15 *top word* e dei 15 *top hashtag* del topic 7.

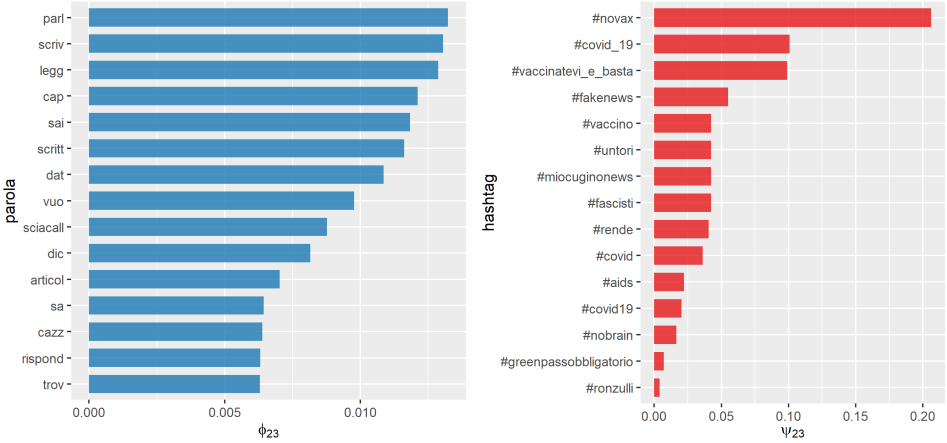


Figura 6.14: Lista delle 15 *top word* e dei 15 *top hashtag* del topic 23.

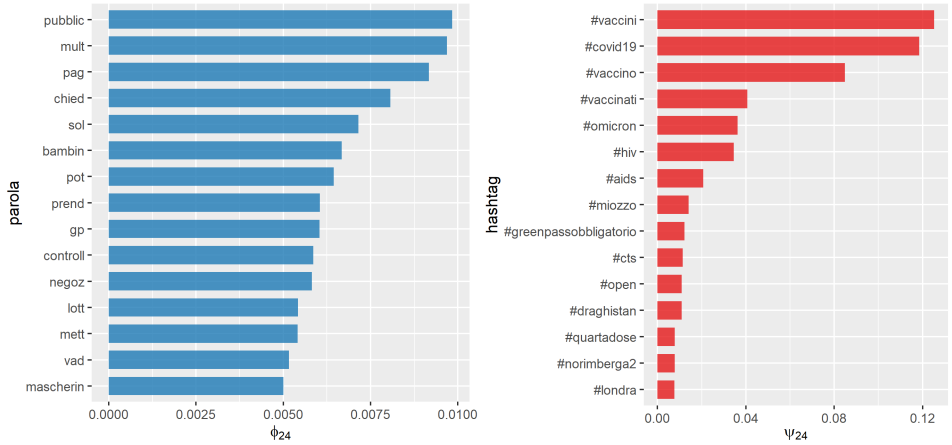


Figura 6.15: Lista delle 15 *top word* e dei 15 *top hashtag* del topic 24.

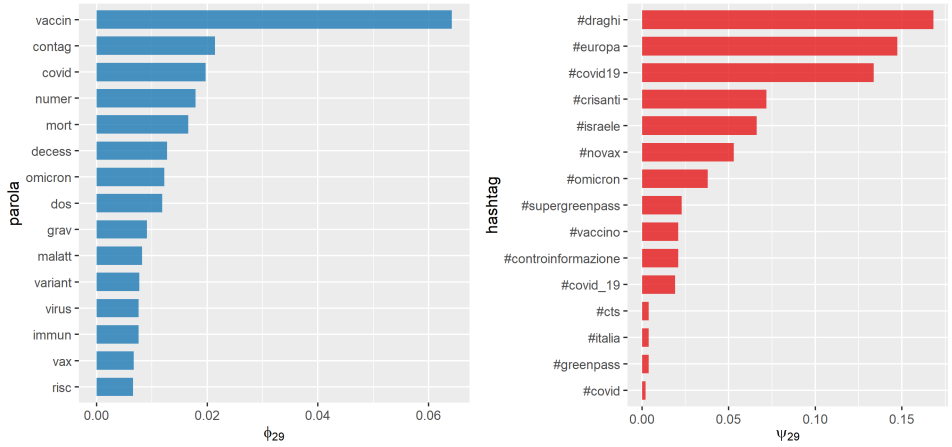


Figura 6.16: Lista delle 15 *top word* e dei 15 *top hashtag* del topic 29.

6.4.5 Parole di Sottofondo e Hashtag Globali

La probabilità che una parola sia generata a partire da un topic è $\pi^V = 0.6776$; si ha quindi che un numero non trascurabile di parole all'interno dei *tweet* è di sottofondo. In Figura 6.19 si può notare che otto parole su quindici sono presenti sia nelle prime 15 *top word* della distribuzione sulle parole della collezione, $\boldsymbol{\phi}^{\mathbf{C}}$, sia nelle prime 15 *top word* della distribuzione sulle parole delle parole di sottofondo, $\boldsymbol{\phi}^{\mathbf{B}}$.

Al contrario, la probabilità che un *hashtag* sia generato a partire da un topic è $\pi^H = 0.9962$; si ha quindi che un numero estremamente basso di *hashtag* all'interno dei *tweet* è *globale*. In questa collezione non esistono

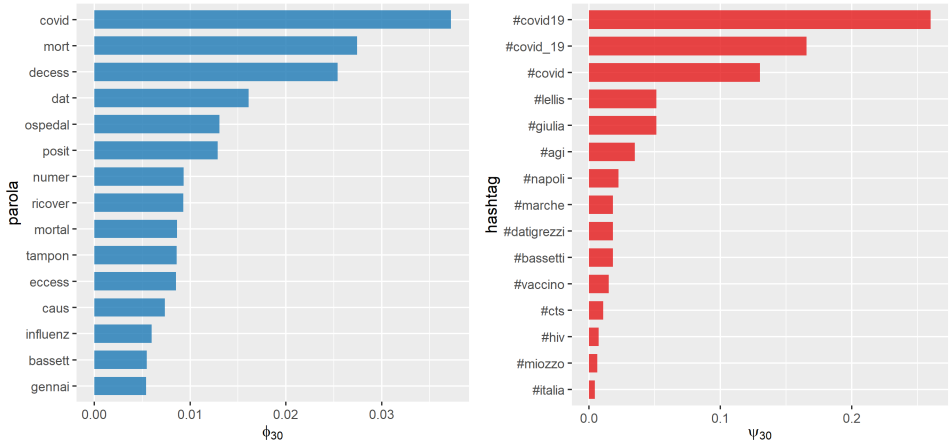


Figura 6.17: Lista delle 15 *top word* e dei 15 *top hashtag* del topic 30.

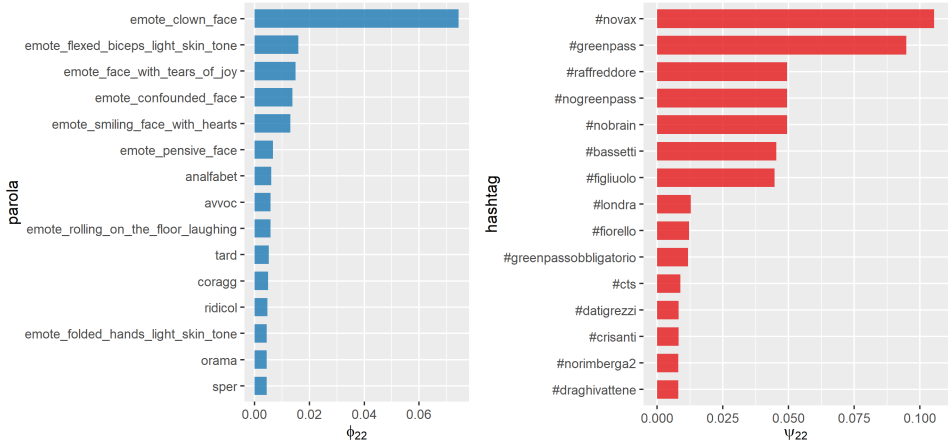


Figura 6.18: Lista delle 15 *top word* e dei 15 *top hashtag* del topic 22.

quindi *hashtag* talmente generali e diffusi da poter essere considerati propri di tutti i topic; la loro estrema rarità porta a una distribuzione sugli *hashtag* degli *hashtag globali*, $\boldsymbol{\psi}^{\mathbf{B}}$, essenzialmente casuale che non risulta simile alla distribuzione sugli *hashtag* della collezione, $\boldsymbol{\psi}^{\mathbf{C}}$. In Figura 6.20 si può notare che tutti i *top hashtag* dal nono in poi della distribuzione sugli *hashtag* degli *hashtag globali* hanno lo stesso peso: si ha ciò poiché quei particolari *hashtag* sono sempre stati generati a partire da un topic nella collezione presa in esame.

Infine, confrontando le distanze tra distribuzioni con la *divergenza di Jensen-Shannon*, si conferma che le due distribuzioni sulle parole siano molto

più simili rispetto alle loro controparti con gli *hashtag*:

$$JSD(\boldsymbol{\phi}^{\mathbf{C}}, \boldsymbol{\phi}^{\mathbf{B}}) = 0.1229$$
$$JSD(\boldsymbol{\psi}^{\mathbf{C}}, \boldsymbol{\psi}^{\mathbf{B}}) = 0.1560$$

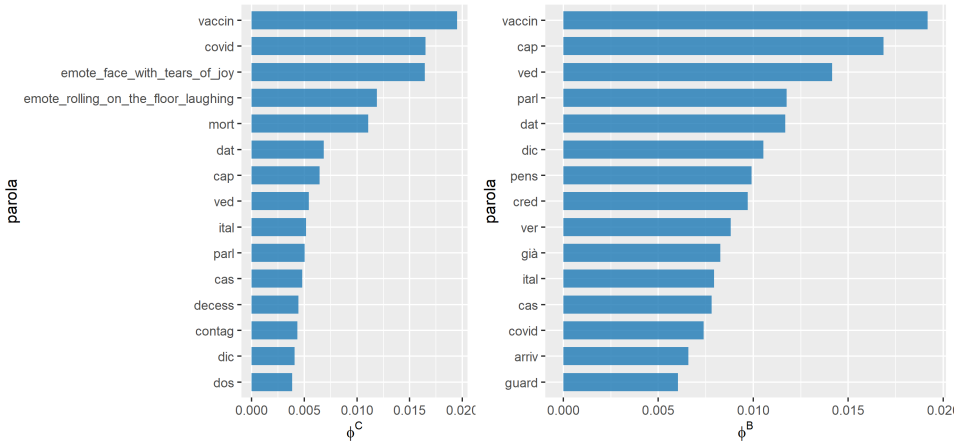


Figura 6.19: Lista delle 15 *top word* della distribuzione sulle parole della collezione, $\boldsymbol{\phi}^{\mathbf{C}}$, e della distribuzione sulle parole delle parole di sottofondo, $\boldsymbol{\phi}^{\mathbf{B}}$.

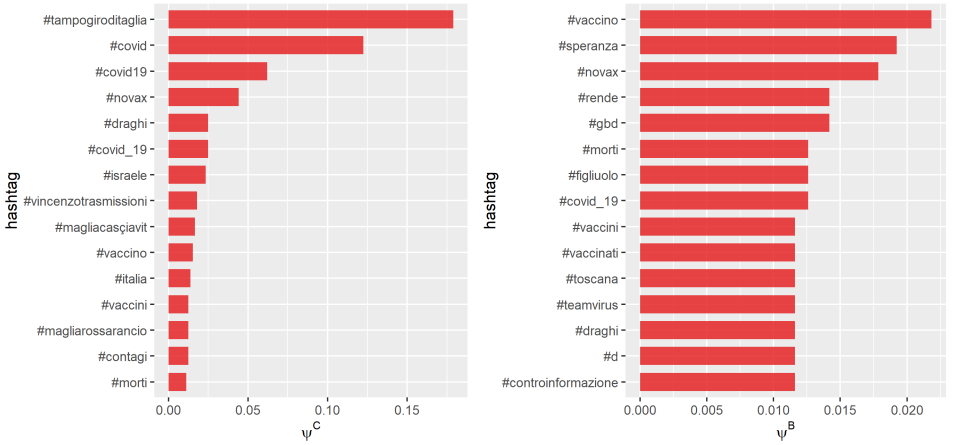


Figura 6.20: Lista dei primi 15 *top hashtag* della distribuzione sugli *hashtag* della collezione, $\boldsymbol{\psi}^{\mathbf{C}}$, e della distribuzione sugli *hashtag* degli *hashtag globali*, $\boldsymbol{\psi}^{\mathbf{B}}$.

Capitolo 7

Conclusioni e Sviluppi Futuri

In questa tesi è stato proposto un nuovo *topic model* che combina le assunzioni di *Latent Dirichlet Allocation*, *Twitter-LDA* e *Hashtag-LDA* e le fa coesistere in modo da rendere possibile un'analisi più accurata e approfondita di collezioni contenenti sia documenti concisi e semplici sia documenti lunghi ed elaborati. Più nello specifico, il *processo generativo* del modello proposto è costruito in modo tale da contenere al suo interno sia quello della *Latent Dirichlet Allocation* sia quello dell'*Hashtag-LDA*, aggiungendo inoltre una distribuzione per le parole di sottofondo in modo da rendere strutturalmente equivalente la generazione delle parole e quella degli *hashtag*. L'aspetto innovativo è la doppia rappresentazione dei documenti, sia come mistura di topic sia come singolo topic, e l'introduzione di una variabile indicatrice che determina quando è ragionevole utilizzare la prima e quando la seconda rappresentazione.

Gli esperimenti sulla collezione di *tweet* forniscono risultati incoraggianti: in termini di *topic coherence*, il modello proposto sembra essere essere nettamente preferibile rispetto alla *Latent Dirichlet Allocation* e *Hashtag-LDA*, e, anche al variare del numero di topic, ottiene risultati quasi coincidenti con *Twitter-LDA*; tuttavia, è importante notare che il modello proposto permette di analizzare più a fondo le relazioni statistiche tra gli elementi testuali contenuti nei testi poiché considera una distinzione tra parole e *hashtag* non presente in *Twitter-LDA*. In particolare, si osserva che l'identificazione della tematica di alcuni topic è possibile solo grazie alla loro doppia rappresentazione, come distribuzione sulle parole e come distribuzione sugli *hashtag*.

Inoltre, osservando le metriche di *topic coherence*, la presenza di parole di sottofondo sembra influenzare positivamente la qualità dei topic identificati.

Sviluppi Futuri

Il lavoro svolto in questa tesi può essere ulteriormente sviluppato seguendo diverse direzioni, sia teoriche sia puramente applicative; alcuni dei possibili sviluppi futuri sono riportati di seguito.

Efficienza Allo stato attuale, la più grande limitazione del modello proposto in questa tesi è il tempo necessario per effettuare l'inferenza approssimata a posteriori. Una possibile direzione è quella di realizzare un'implementazione in un altro linguaggio di programmazione, ad esempio il *C*. Alternativamente, si potrebbe valutare l'utilizzo di altri algoritmi per l'inferenza, come ad esempio *Variational Inference*.

Estensione non parametrica Il modello proposto in questa tesi è un *topic model* parametrico in cui è necessario fissare a priori il numero di topic: potrebbe essere interessante valutare l'utilizzo di *Hierarchical Dirichlet Process* (Teh et al., 2006) al posto delle distribuzioni di *Dirichlet* per far scegliere al modello stesso il numero di topic ottimale.

Dizionari multipli Distinguere parole e *hashtag* significa assumere che gli elementi del testo di un documento non provengano tutti dalla stesso vocabolario, ma da vocabolari diversi definiti secondo un criterio fissato. In questa tesi ci si limita a distinguere tra parole e *hashtag*, tuttavia si potrebbero, ad esempio, distinguere parole, *hashtag*, *emoji*, nomi propri di persona, nomi propri di città, e così via. Si noti che definire un vocabolario significa implicitamente che si è interessati a valutare l'importanza di quei termini in maniera indipendente da quelli presenti in altri vocabolari.

Utenti e non solo Riprendendo l'intuizione introdotta nella sottosezione 2.3.2, l'utente che ha scritto ogni documento può essere visto come un'informazione aggiuntiva legata al documento attraverso il *tagging*. Sostituendo quindi il *tag* "Utente:" con una qualsiasi altra informazione aggiuntiva disponibile, è possibile valutare come questa caratteristica influenza il contenuto

dei documenti. Ad esempio, si potrebbe essere interessati alla distribuzione sui topic di diversi gruppi di utenti, di documenti scritti in diversi archi temporali, di documenti contenenti determinate parole chiave, e così via.

Raccomandazione di Hashtag *Hashtag-LDA* (F. Zhao et al., 2016) nasce non come strumento per identificare le tematiche di una collezione, ma come metodo di raccomandazione di *hashtag*: dato lo stretto legame tra questo modello e quello proposto in questa tesi, potrebbe essere interessante effettuare ulteriori esperimenti in cui il modello proposto viene valutato come metodo di raccomandazione.

Oltre i microblog Come suggerisce il titolo stesso della tesi, *Un algoritmo di topic modeling per microblog*, il modello proposto è nato ed è stato sviluppato avendo come obiettivo l'estrazione di informazione da testi provenienti da *microblog*. Tuttavia è ragionevole valutare un'applicazione anche a testi provenienti da fonti alternative. Ad esempio, potrebbe essere applicato alle discussioni pubblicate in un *forum*, ai post e ai commenti pubblicati in un *blog*, ai post pubblicati su *Reddit*, e così via. Più in generale, è ragionevole applicare il modello a qualsiasi collezione formata sia da documenti semplici e concisi sia documenti lunghi ed elaborati.

Appendice A

Distribuzioni di Probabilità

Si introducono le distribuzioni di probabilità utilizzate nella tesi; infine, nell'ultima sezione si descrivono le principali differenze tra la distribuzione categoriale e la distribuzione multinomiale, inoltre si spiega in quali situazioni è necessario utilizzare la prima e quando la seconda. Le notazione è tratta da Pace e Salvan, 2001, mentre le definizioni da Bishop, 2006 e Murphy, 2012.

A.1 Distribuzione Beta

La distribuzione di Y si dice Beta con parametri $\alpha > 0$ e $\beta > 0$, e si scrive sinteticamente $Y \sim Beta(\alpha, \beta)$, se Y è continua con supporto $S_Y = [0, 1]$ e funzione di probabilità, per $y \in S_Y$,

$$\begin{aligned} p_Y(y; \alpha, \beta) &= \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \end{aligned}$$

A.2 Distribuzione di Dirichlet

La distribuzione di $\mathbf{Y} = (Y_1, \ldots, Y_K)$ si dice di Dirichlet di ordine $K \geq 2$ con parametro $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\alpha_k > 0$ per $k = 1, \ldots, K$, e si scrive sinteticamente $\mathbf{Y} \sim Dir_K(\boldsymbol{\alpha})$, se $\mathbf{Y}$ ha supporto sul simplesso $(K-1)$-dimensionale

$$S_Y = \left\{ \mathbf{y} = (y_1, \ldots, y_K) \in (0,1)^K : \sum_{k=1}^{K} y_k = 1 \right\}$$

e funzione di probabilità, per $\mathbf{y} \in S_Y$,

$$p_Y(\mathbf{y}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} y_1^{\alpha_1 - 1} \cdots y_K^{\alpha_K - 1}$$

La distribuzione di Dirichlet è l'estensione multivariata della distribuzione Beta, infatti è anche detta distribuzione Beta multivariata (*multivariate beta distribution, MBD*). Se $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$, ovvero $\alpha_1 = \ldots = \alpha_K = \alpha$, allora $\mathbf{Y}$ segue una distribuzione di Dirichlet simmetrica di parametro α, si scrive sinteticamente $\mathbf{Y} \sim Dir_K(\alpha)$, $\alpha > 0$, e la funzione di probabilità diventa

$$p_Y(\mathbf{y}; \alpha) = \frac{\Gamma(K\alpha)}{K\Gamma(\alpha)} (y_1 \cdots y_K)^{\alpha - 1}$$

A.3 Distribuzione di Bernoulli

La distribuzione di Y si dice di Bernoulli con parametro $\pi \in (0, 1)$, e si scrive sinteticamente $Y \sim Bern(\pi)$, se Y è discreta con supporto $S_Y = \{0, 1\}$ e funzione di probabilità, per $y \in S_Y$,

$$p_Y(y; \pi) = \pi^y (1 - \pi)^{(1-y)}$$

A.4 Distribuzione Categoriale

La distribuzione di Y si dice categoriale con vettore di probabilità $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)$, $\pi_i \in (0, 1)$ per $i = 1, \ldots, d$, e si scrive sinteticamente $Y \sim Cat(\boldsymbol{\pi})$, se Y è discreta con supporto $S_Y = \{1, \ldots, d\}$ e funzione di probabilità, per $y \in S_Y$,

$$p_Y(y; \boldsymbol{\pi}) = \pi_1^{\mathbb{1}_{y=1}} \cdots \pi_d^{\mathbb{1}_{y=d}}$$

A.5 Distribuzione Multinomiale

La distribuzione di $\mathbf{Y} = (Y_1, \ldots, Y_d)$ si dice multinomiale con indice $n \in \mathbb{N}$ e vettore di probabilità $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)$, $\pi_i \in (0, 1)$ per $i = 1, \ldots, d$, e si scrive sinteticamente $\mathbf{Y} \sim Mult(n, \boldsymbol{\pi})$, se $\mathbf{Y}$ è discreta con supporto

$$S_Y = \left\{ \mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{N}^d : \sum_{i=1}^d y_i = n \right\}$$

e funzione di probabilità, per $\mathbf{y} \in S_Y$,

$$p_Y(\mathbf{y}; \boldsymbol{\pi}) = \frac{n!}{y_1! \cdots y_d!} \pi_1^{y_1} \cdots \pi_d^{y_d}$$

A.6 Confronto tra Distribuzione Categoriale e Distribuzione Multinomiale

Si può facilmente dimostrare che una distribuzione categoriale e una distribuzione multinomiale con indice $n = 1$ descrivono esattamente la stessa situazione, se si considera lo stesso vettore di probabilità $\boldsymbol{\pi}$, utilizzando però una diversa rappresentazione della variabile aleatoria Y. In particolare, in entrambi i casi Y può assumere d valori ma nel primo caso Y è uno scalare che assume valori in {1,. . . ,d}, e nel secondo $\mathbf{Y}$ è un vettore d-dimensionale con un unico valore non nullo e pari a 1.

Si consideri ad esempio il risultato del lancio di un dado a 3 facce: si ha $d = 3$ e il risultato 2 può essere rappresentato come $Y = 2$ oppure $\mathbf{Y} = (0, 1, 0)$. Questa corrispondenza uno a uno può essere ricavata a partire dalla densità delle due distribuzioni; siano $d = 3$, $Y \sim Cat(\boldsymbol{\pi})$ e $\mathbf{Y} \sim Mult(\boldsymbol{\pi})$, allora vale:

$$p(y; \boldsymbol{\pi}) = \pi_1^{\mathbb{1}_{y=1}} \pi_2^{\mathbb{1}_{y=2}} \pi_3^{\mathbb{1}_{y=3}} = \prod_{i=1}^{3} \pi_i^{\mathbb{1}_{y=i}} = \begin{cases} \pi_1 & \text{se } y = 1 \\ \pi_2 & \text{se } y = 2 \\ \pi_3 & \text{se } y = 3 \end{cases}$$

$$p(\mathbf{y}; \boldsymbol{\pi}) = \pi_1^{y_1} \pi_2^{y_2} \pi_3^{y_3} = \prod_{i=1}^{3} \pi_i^{y_i} = \begin{cases} \pi_1 & \text{se } \mathbf{y} = (1, 0, 0) \\ \pi_2 & \text{se } \mathbf{y} = (0, 1, 0) \\ \pi_3 & \text{se } \mathbf{y} = (0, 0, 1) \end{cases}$$

Le parole, gli *hashtag* e i topic possono essere rappresentati utilizzando una delle due rappresentazioni e, in base alla scelta, diventa necessario utilizzare la distribuzione categoriale oppure la distribuzione multinomiale: a livello pratico cambia solo la notazione. In questa tesi, parole, *hashtag* e topic sono rappresentati come scalari, quindi seguono delle distribuzioni categoriali.

Appendice B

Logaritmo delle Distribuzioni Congiunte

B.1 Latent Dirichlet Allocation

Il logaritmo della distribuzione congiunta delle variabili osservate e latenti della *Latent Dirichlet Allocation*, indicata con ℓ_{LDA}, è proporzionale a:

$$\begin{aligned}\ell_{LDA} &= \log p(\mathbf{z}^V, \mathbf{w}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\phi}_{1:T} | \alpha, \beta^V) \\ &\propto \sum_{d=1}^{D}\sum_{t=1}^{T}(\alpha_t + n^t_{z^V_d} - 1)\log\theta_{d,t} + \sum_{t=1}^{T}\sum_{v=1}^{V}(\beta^V_v + n^{v,t} - 1)\log\phi_{t,v}\end{aligned}$$

B.2 Twitter-LDA

Il logaritmo della distribuzione congiunta delle variabili osservate e latenti di *Twitter-LDA*, indicata con ℓ_{T-LDA}, è proporzionale a:

$$\begin{aligned}\ell_{TLDA} =& \log p(\mathbf{w}, \mathbf{y}^V, \mathbf{z}^*, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\phi}^{\mathcal{B}}, \pi^V | \alpha^*, \beta^V, b^V_1, b^V_2) \\ =& (b^V_1 + n^1_{y^V} - 1)\log\pi^V + (b^V_2 + N - n^1_{y^V} - 1)\log(1-\pi^V) \\ &+ \sum_{u=1}^{U}\sum_{t=1}^{T}(\alpha^*_t + n^t_{z^*_u} - 1)\log\theta^*_{u,t} \\ &+ \sum_{v=1}^{V}(\beta^V_v + n^v_{y^V=0} - 1)\log\phi^{\mathcal{B}}_v \\ &+ \sum_{t=1}^{T}\sum_{v=1}^{V}(\beta^V_v + n^{v,t}_{y^V=1} - 1)\log\phi_{t,v}\end{aligned}$$

B.3 Hashtag-LDA

Il logaritmo della distribuzione congiunta delle variabili osservate e latenti di *Hashtag-LDA*, indicata con ℓ_{H-LDA}, è proporzionale a:

$$
\begin{aligned}
\ell_{HLDA} &= \log p(\mathbf{w}, \mathbf{h}, \mathbf{y}^H, \mathbf{z}^*, \boldsymbol{\theta}^*_{1:U}, \boldsymbol{\phi}_{1:T}, \boldsymbol{\psi}_{1:T}\boldsymbol{\phi}^{\mathcal{B}}, \pi^H | \alpha^*, \beta^V, \beta^H, b_1^H, b_2^H) \\
&\propto (b_1^H + n^1_{y^H} - 1)\log \pi^H + (b_2^H + L - n^1_{y^H} - 1)\log(1-\pi^H) \\
&\quad + \sum_{u=1}^{U}\sum_{t=1}^{T}(\alpha_t^* + n^t_{z_u^*} - 1)\log\theta^*_{u,t} \\
&\quad + \sum_{t=1}^{T}\sum_{v=1}^{V}(\beta_v^V + n^{v,t} - 1)\log\phi_{t,v} \\
&\quad + \sum_{h=1}^{H}(\beta_h^H + n^h_{y^H=0} - 1)\log\psi_h^{\mathcal{B}} \\
&\quad + \sum_{t=1}^{T}\sum_{h=1}^{H}(\beta_h^H + n^{h,t}_{y^H=1} - 1)\log\phi_{t,h}
\end{aligned}
$$

B.4 Modello proposto

Il logaritmo della distribuzione congiunta delle variabili osservate e latenti del modello proposto, indicata con ℓ_{M-LDA}, è proporzionale a:

$$
\begin{aligned}
\ell_{MLDA} &= \log p(\mathbf{lat}, \mathbf{par} | \boldsymbol{\alpha}^*, \boldsymbol{\alpha}, \alpha_0, \boldsymbol{\beta}^V, \boldsymbol{\beta}^H, \mathbf{b}) \\
&\propto (b_1^\delta + n^1_\lambda - 1)\log\delta + (b_2^\delta + DT - n^1_\lambda - 1)\log(1-\delta) \\
&\quad + (b_1^H + n^1_{y^H} - 1)\log\pi^H + (b_2^H + L - n^1_{y^H} - 1)\log(1-\pi^H) \\
&\quad + (b_1^H + n^1_{y^H} - 1)\log\pi^H + (b_2^H + L - n^1_{y^H} - 1)\log(1-\pi^H) \\
&\quad + \sum_{u=1}^{U}\left[(b_1^T + n^1_{x_u} - 1)\log\pi_u^T + (b_2^T + D_u - n^1_{x_u} - 1)\log(1-\pi_u^T)\right] \\
&\quad + \sum_{u=1}^{U}\sum_{t=1}^{T}(\alpha_t^* + n^t_{z_u^*} - 1)\log\theta^*_{u,t} + \\
&\quad + \sum_{u=1}^{U}\sum_{d=1}^{D_u}\log\left(\frac{\Gamma(T\alpha_0 + \sum_{t=1}^{T}\lambda_{ud,t}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_0 + \lambda_{ud,t}\alpha_t)}\right) \\
&\quad + \sum_{u=1}^{U}\sum_{d=1}^{D_u}\sum_{t=1}^{T}(\alpha_0 + \lambda_{ud,t}\alpha_t + n^t_{z^V_{ud}, z^H_{ud}} - 1)\log\theta_{ud,t}
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{v=1}^{V} (\beta_v^V + n_{y^V=0}^v - 1) \log \phi_v^{\mathcal{B}} \\
&+ \sum_{t=1}^{T} \sum_{v=1}^{V} (\beta_v^V + n_{y^V=1}^{v,t} - 1) \log \phi_{t,v} \\
&+ \sum_{h=1}^{H} (\beta_h^H + n_{y^H=0}^h - 1) \log \psi_h^{\mathcal{B}} \\
&+ \sum_{t=1}^{T} \sum_{h=1}^{H} (\beta_h^H + n_{y^H=1}^{h,t} - 1) \log \phi_{t,h}
\end{aligned}
$$

Appendice C

Codice R

C.1 Librerie

Le librerie di *R* utilizzate in questa tesi sono:

- `dplyr` Wickham et al., 2021
- `ggplot2` Wickham, 2016
- `gridExtra` Auguie, 2017
- `latex2exp` Meschiari, 2022
- `Matrix` Bates e Maechler, 2021
- `quanteda` Benoit et al., 2018
- `stringr` Wickham, 2019
- `TextWiller` Solari et al., 2019
- `tidytext` Silge e Robinson, 2016
- `topicdoc` Friedman, 2019
- `topicmodels` Grün e Hornik, 2011
- `wordcloud2` Lang e Chien, 2018

C.2 Collapsed Gibbs Sampler

```
# ------------------------------------------------------------------------ #

CGS_Microblog <- function(w, h, doc_users,
                          alphastar, alpha, betaV, betaH, bV, bH, bdelta, bT,
                          alpha0=10^-7, iterations=300, seed=28, result_folder)
    {
  # ------------------------------------------------------------------------ #
  # Argomenti della funzione:
  #          w : matrice D x Nmax   | n-ma parola del d-mo documento
  #          h : matrice D x Lmax   | l-mo hashtag del d-mo documento
  #  doc_users : vettore D x 1      | autore del d-mo topic
  #  alphastar : vettore TOPICS x 1 | parametro Dirichlet sul simplesso dei
    topic
  #      alpha : vettore TOPICS x 1 | smoothing prior (parametro Dirichlet sul
    simplesso dei topic)
  #      betaV : vettore V x 1      | parametro Dirichlet sul simplesso delle
    parole
  #      betaH : vettore H x 1      | parametro Dirichlet sul simplesso degli
    hashtag
  #         bV : vettore 2x1        | parametro Beta
  #         bH : vettore 2x1        | parametro Beta
  #     bdelta : vettore 2x1        | parametro Beta
  #         bT : vettore 2x1        | parametro Beta
  #     alpha0 : reale positivo     | weak smoothing prior (10^-7)
  # iterations : intero             | numero di stati della catena da campionare
  #       seed : intero             | seme per rendere i risultati replicabili
  # ------------------------------------------------------------------------ #
  cat("\n", as.character(Sys.time()), " Operazioni preliminari.", sep="")
  # Importo librerie
  require(Matrix)
  require(stringr)
  # Fisso il seme
  set.seed(seed)
  # Creo cartella in cui salvare gli stati della catena
  result_folder <- file.path(getwd(), "results", result_folder)
  if(!dir.exists(result_folder)) {
    dir.create(result_folder)
    dir.create(file.path(result_folder, "x"))
    dir.create(file.path(result_folder, "zstar"))
    dir.create(file.path(result_folder, "lambda"))
    dir.create(file.path(result_folder, "yV"))
    dir.create(file.path(result_folder, "zV"))
    dir.create(file.path(result_folder, "yH"))
    dir.create(file.path(result_folder, "zH"))
  } else {
    cat("\nLa cartella '",result_folder,"' esiste gia': selezionare un altro
     valore per la variabile 'result_folder'.\n", sep="")
```

```
42    return(NULL)
43  }
44  # Definisco alcune quantita' utili
45  TOPICS <- length(alpha)
46  U <- length(unique(doc_users))
47  D <- nrow(w)
48  V <- length(betaV)
49  H <- length(betaH)
50  N <- sum(w>0)
51  L <- sum(h>0)
52  betaV_sum <- sum(betaV)
53  betaH_sum <- sum(betaH)
54  # ------------------------------------------------------------------------ #
55  # GENERAZIONE MATRICI DI CONTEGGI
56  cat("\n", as.character(Sys.time()), " Creazione delle matrici di conteggi.",
      sep="")
57  # Creo le matrici dei conteggi
58  X1 <- rep(0, U)
59  WY1ZX <- matrix(0, nrow=V, ncol=TOPICS)
60  HY1ZX <- matrix(0, nrow=H, ncol=TOPICS)
61  Zstar <- matrix(0, nrow=U, ncol=TOPICS)
62  LAMBDA1 <- 0
63  Yv1 <- 0
64  WY0 <- rep(0, V)
65  Yh1 <- 0
66  HY0 <- rep(0, H)
67  Z <- matrix(0, nrow=D, ncol=TOPICS)
68  WY1Z <- list(); for (d in 1:D) WY1Z[[d]] <- Matrix(0, nrow=V, ncol=TOPICS,
      sparse=T)
69  HY1Z <- list(); for (d in 1:D) HY1Z[[d]] <- Matrix(0, nrow=H, ncol=TOPICS,
      sparse=T)
70  # FINE GENERAZIONE MATRICI DI CONTEGGI
71  # ------------------------------------------------------------------------ #
72  # GENERAZIONE STATO INIZIALE
73  cat("\n", as.character(Sys.time()), " Generazione dello stato iniziale: ", sep
      ="")
74  # Genero le var. dello stato iniziale e in contemporanea aggiorno i conteggi
75  x <- rep(0, D)
76  zstar <- rep(0, D)
77  lambda <- matrix(0, nrow=D, ncol=TOPICS)
78  yV <- zV <- matrix(0, nrow=D, ncol=ncol(w))
79  yH <- zH <- matrix(0, nrow=D, ncol=ncol(h))
80  # ------------------------------------------------------------------------ #
81  for (d in 1:D) {
82    if (d %in% round(quantile(1:D,(1:10)/10))) cat("-")
83    u <- doc_users[d]   # autore del documento
84    Nd_range <- if(sum(w[d,]>0)>0) 1:sum(w[d,]>0) else c()  # indici delle
       parole nel documento
85    Ld_range <- if(sum(h[d,]>0)>0) 1:sum(h[d,]>0) else c()  # indici degli
       hashtag nel documento
```

```
    # estraggo valori
    x[d] <- rbinom(1, size=1, p=0.5)
    zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=alphastar/sum(
     alphastar))
    lambda[d,] <- rbinom(TOPICS, size=1, p=bdelta[1]/sum(bdelta))
    # aggiorno i conteggi
    X1[u] <- X1[u] + x[d]
    Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
    LAMBDA1 <- LAMBDA1 + sum(lambda[d,])
    for (n in Nd_range) {
      # estraggo valori
      yV[d,n] <- rbinom(1, size=1, p=bV[1]/sum(bV))
      zV[d,n] <- sample.int(TOPICS, size=1, replace=T, prob=alpha/sum(alpha))
      # aggiorno i conteggi
      Z[d,zV[d,n]] <- Z[d,zV[d,n]] + 1
      if (yV[d,n]==1) {
        if (x[d]==1) {
          WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + 1
        } else {
          WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
        }
        Yv1 <- Yv1 + 1
        WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
      } else {
        WY0[w[d,n]] <- WY0[w[d,n]] + 1
      }
    }
    for (l in Ld_range) {
      # estraggo valori
      yH[d,l] <- rbinom(1, size=1, p=bH[1]/sum(bH))
      zH[d,l] <- sample.int(TOPICS, size=1, replace=T, prob=alpha/sum(alpha))
      # aggiorno i conteggi
      Z[d,zH[d,l]] <- Z[d,zH[d,l]] + 1
      if (yH[d,l]==1) {
        if (x[d]==1) {
          HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] + 1
        } else {
          HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
        }
        Yh1 <- Yh1 + 1
        HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
      } else {
        HY0[h[d,l]] <- HY0[h[d,l]] + 1
      }
    }
  }
  # FINE GENERAZIONE STATO INIZIALE
  # ------------------------------------------------------------------------ #
  # COLLAPSED GIBBS SAMPLING (CGS)
```

```
134  cat("\n", as.character(Sys.time()), " Collapsed Gibbs Sampler (", iterations,
       " iterazioni)", sep="")
135  for (m in 1:iterations) {
136    # ------------------------------------------------------------------------ #
137    # ITERAZIONE CGS
138    cat("\n", as.character(Sys.time()), "  - iterazione ", str_pad(m, 3, pad = "
       "), ": ", sep="")
139    for (d in 1:D) {
140      # ------------------------------------------------------------------------ #
141      # AGGIORNAMENTO DOCUMENTO
142      if (d %in% round(quantile(1:D,(1:20)/20))) cat("-")
143      # definisco alcune quantita' utili
144      u <- doc_users[d]   # autore del documento
145      Nd <- sum(w[d,]>0)
146      if(Nd>0) {
147        Nd_range <-  1:Nd                        # indici delle parole nel
       documento
148        unique_words <- unique(w[d,Nd_range])    # parole distinte nel documento
149      } else {
150        Nd_range <- unique_words <- c()
151      }
152      Ld <- sum(h[d,]>0)
153      if(Ld>0) {
154        Ld_range <- 1:Ld                         # indici degli hashtag nel
       documento
155        unique_hashtags <- unique(h[d,Ld_range]) # hashtag distinti nel
       documento
156      } else {
157        Ld_range <- unique_hashtags <- c()
158      }
159      # ------------------------------------------------------------------------ #
160      # AGGIORNAMENTO x[d]
161      # rimuovo x[d] dai conteggi
162      X1[u] <- X1[u] - x[d]
163      if (x[d]==1) {
164        for (n in Nd_range) WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] - yV[
       d,n]
165        for (l in Ld_range) HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] - yH[
       d,l]
166      } else {
167        for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] -
       yV[d,n]
168        for (l in Ld_range) HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] -
       yH[d,l]
169      }
170      # calcolo la prob. di avere x[d]=1 e x[d]=0
171      p_x1 <- bT[1] + X1[u]
172      p_x0 <- bT[2] + D-1 - X1[u]
173      qmax  <- sum(WY1Z[[d]])
174      if (qmax>0) {
```

```
175         p_x0_num <- c(); p_x0_den <- betaV_sum + sum(WY1ZX[,zstar[d]]) + 0:(qmax
    -1)
176         for (t in unique(c(zstar[d],zV[d,Nd_range]))) {
177           qmax_t <- sum(WY1Z[[d]][,t])
178           if (qmax_t>0) {
179             p_x1_num <- c(); p_x1_den <- betaV_sum + sum(WY1ZX[,t]) + 0:(qmax_t-
    1)
180           }
181           for (w_dn in unique_words) {
182             # determino il numero di termini delle produttorie in q
183             qmax_v  <- sum(WY1Z[[d]][w_dn,])
184             # aggiorno probabilita'
185             if (qmax_v>0) {
186               if (t==zstar[d]) p_x0_num <- c(p_x0_num, betaV[w_dn] + WY1ZX[w_dn,
    t] + 0:(qmax_v-1))
187               if (WY1Z[[d]][w_dn,t]>0) p_x1_num <- c(p_x1_num, betaV[w_dn] +
    WY1ZX[w_dn,t] + 0:(WY1Z[[d]][w_dn,t]-1))
188             }
189           }
190           if (qmax_t>0) p_x1 <- p_x1 * prod(p_x1_num/p_x1_den)
191         }
192         p_x0 <- p_x0 * prod(p_x0_num/p_x0_den)
193       }
194       qmax  <- sum(HY1Z[[d]])
195       if (qmax>0) {
196         p_x0_num <- c(); p_x0_den <- betaH_sum + sum(HY1ZX[,zstar[d]]) + 0:(qmax
    -1)
197         for (t in unique(c(zstar[d],zH[d,Ld_range]))) {
198           qmax_t <- sum(HY1Z[[d]][,t])
199           if (qmax_t>0) {
200             p_x1_num <- c(); p_x1_den <- betaH_sum + sum(HY1ZX[,t]) + 0:(qmax_t-
    1)
201           }
202           for (h_dl in unique_hashtags) {
203             # determino il numero di termini delle produttorie in q
204             qmax_h  <- sum(HY1Z[[d]][h_dl,])
205             # aggiorno probabilita'
206             if (qmax_h>0) {
207               if (t==zstar[d]) p_x0_num <- c(p_x0_num, betaH[h_dl] + HY1ZX[h_dl,
    t] + 0:(qmax_h-1))
208               if (HY1Z[[d]][h_dl,t]>0) p_x1_num <- c(p_x1_num, betaH[h_dl] +
    HY1ZX[h_dl,t] + 0:(HY1Z[[d]][h_dl,t]-1))
209             }
210           }
211           if (qmax_t>0) p_x1 <- p_x1 * prod(p_x1_num/p_x1_den)
212         }
213         p_x0 <- p_x0 * prod(p_x0_num/p_x0_den)
214       }
215       # aggiorno x[d]
216       x[d] <- sample(c(0,1), size=1, prob=c(p_x0,p_x1)/sum(p_x0,p_x1))
```

```
217       # aggiorno i conteggi
218       X1[u] <- X1[u] + x[d]
219       if (x[d]==1) {
220         for (n in Nd_range) WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + yV[
      d,n]
221         for (l in Ld_range) HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] + yH[
      d,l]
222       } else {
223         for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] +
      yV[d,n]
224         for (l in Ld_range) HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] +
      yH[d,l]
225       }
226       # FINE AGGIORNAMENTO x[d]
227       # ------------------------------------------------------------------ #
228       # AGGIORNAMENTO zstar[d]
229       if (x[d]==1) {
230         # rimuovo zstar[d] dai conteggi
231         Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] - 1
232         for (n in Nd_range) WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] - yV[
      d,n]
233         for (l in Ld_range) HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] - yH[
      d,l]
234         # calcolo la prob. di avere zstar[d]=t
235         p_zstar <- (alphastar + Zstar[u,])
236         # aggiorno zstar[d]
237         zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=p_zstar/sum(p_
      zstar))
238         # aggiorno i conteggi
239         Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
240         for (n in Nd_range) WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + yV[
      d,n]
241         for (l in Ld_range) HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] + yH[
      d,l]
242       } else {
243         # rimuovo zstar[d] dai conteggi
244         Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] - 1
245         for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] -
      yV[d,n]
246         for (l in Ld_range) HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] -
      yH[d,l]
247         # calcolo la prob. di avere zstar[d]=t
248         p_zstar <- (alphastar + Zstar[u,])
249         for (t in 1:TOPICS) {
250           qmax  <- sum(WY1Z[[d]])
251           if (qmax>0) {
252             p_num <- c(); p_den <- betaV_sum + sum(WY1ZX[,t]) + 0:(qmax-1)
253             for (w_dn in unique_words) {
254               # determino il numero di termini della produttoria in q
255               qmax_v  <- sum(WY1Z[[d]][w_dn,])
```

```
256             if (qmax_v>0) p_num <- c(p_num, betaV[w_dn] + WY1ZX[w_dn,t] + 0:(
    qmax_v-1))
257           }
258           p_zstar[t] <- p_zstar[t] * prod(p_num/p_den)
259         }
260         qmax  <- sum(HY1Z[[d]])
261         if (qmax>0) {
262           p_num <- c(); p_den <-betaH_sum + sum(HY1ZX[,t]) + 0:(qmax-1)
263           for (h_dl in unique_hashtags) {
264             # determino il numero di termini della produttoria in q
265             qmax_h  <- sum(HY1Z[[d]][h_dl,])
266             # aggiorno probabilita'
267             if (qmax_h>0) p_num <- c(p_num, betaH[h_dl] + HY1ZX[h_dl,t] + 0:(
    qmax_h-1))
268           }
269           p_zstar[t] <- p_zstar[t] * prod(p_num/p_den)
270         }
271       }
272       # aggiorno zstar[d]
273       zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=p_zstar/sum(p_
    zstar))
274       # aggiorno i conteggi
275       Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
276       for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] +
    yV[d,n]
277       for (l in Ld_range) HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] +
    yH[d,l]
278      }
279      # FINE AGGIORNAMENTO zstar[d]
280      # ------------------------------------------------------------------- #
281      # AGGIORNAMENTO lambda[d,]
282      for (t in 1:T) {
283       # ------------------------------------------------------------------- #
284       # AGGIORNAMENTO lambda[d,t]
285       # rimuovo lambda[d,t] dai conteggi
286       LAMBDA1 <- LAMBDA1 - lambda[d,t]
287       # calcolo la prob. di avere lambda[d,t]=0 e lambda[d,t]=1
288       p_lambda0 <- (bdelta[2] + D*TOPICS-1 - LAMBDA1) / prod(TOPICS*alpha0+sum
    (lambda[d,-t]*alpha[-t]) + 0:(Nd+Ld-1))
289       p_lambda1 <- (bdelta[1] + LAMBDA1)              / prod(TOPICS*alpha0+sum
    (lambda[d,]*alpha) + 0:(Nd+Ld-1))
290       if (Z[d,t]>0) {
291         p_lambda0 <- p_lambda0 * prod(alpha0 + 0:(Z[d,t]-1))
292         p_lambda1 <- p_lambda1 * prod(alpha0+alpha[t] + 0:(Z[d,t]-1))
293       }
294       # aggiorno lambda[d,t]
295       lambda[d,t] <- sample(c(0,1), size=1, prob=c(p_lambda0,p_lambda1)/sum(p_
    lambda0,p_lambda1))
296       # aggiorno i conteggi
297       LAMBDA1 <- LAMBDA1 + lambda[d,t]
```

```
298         # FINE AGGIORNAMENTO lambda[d,t]
299         # ------------------------------------------------------------------ #
300       }
301       # FINE AGGIORNAMENTO lambda[d,]
302       # ------------------------------------------------------------------ #
303       for (n in Nd_range) {
304         # ------------------------------------------------------------------ #
305         # AGGIORNAMENTO yV[d,n]
306         if (yV[d,n]==1) {
307           if (x[d]==1) {
308             # rimuovo yV[d,n] dai conteggi
309             WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] - 1
310             Yv1 <- Yv1 - 1
311             WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] - 1
312             # calcolo la prob. di avere yV[d,n]=0 e yV[d,n]=1
313             p_Yv0 <- (bV[2] + N-1 - Yv1) * (betaV[w[d,n]] + WY0[w[d,n]]) / (
    betaV_sum + sum(WY0))
314             p_Yv1 <- (bV[1] + Yv1) * (betaV[w[d,n]] + WY1ZX[w[d,n],zV[d,n]]) / (
    betaV_sum + sum(WY1ZX[,zV[d,n]]))
315             # aggiorno yV[d,n]
316             yV[d,n] <- sample(c(0,1), size=1, prob=c(p_Yv0,p_Yv1)/sum(p_Yv0,p_
    Yv1))
317             # aggiorno i conteggi
318             if (yV[d,n]==1) {
319               WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + 1
320               Yv1 <- Yv1 + 1
321               WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
322             } else {
323               WY0[w[d,n]] <- WY0[w[d,n]] + 1
324             }
325           } else {
326             # rimuovo yV[d,n] dai conteggi
327             WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] - 1
328             Yv1 <- Yv1 - 1
329             WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] - 1
330             # calcolo la prob. di avere yV[d,n]=0 e yV[d,n]=1
331             p_Yv0 <- (bV[2] + N-1 - Yv1) * (betaV[w[d,n]] + WY0[w[d,n]]) / (
    betaV_sum + sum(WY0))
332             p_Yv1 <- (bV[1] + Yv1) * (betaV[w[d,n]] + WY1ZX[w[d,n],zstar[d]]) /
    (betaV_sum + sum(WY1ZX[,zstar[d]]))
333             # aggiorno yV[d,n]
334             yV[d,n] <- sample(c(0,1), size=1, prob=c(p_Yv0,p_Yv1)/sum(p_Yv0,p_
    Yv1))
335             # aggiorno i conteggi
336             if (yV[d,n]==1) {
337               WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
338               Yv1 <- Yv1 + 1
339               WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
340             } else {
341               WY0[w[d,n]] <- WY0[w[d,n]] + 1
```

```
342             }
343           }
344         } else {
345           if (x[d]==1) {
346             # rimuovo yV[d,n] dai conteggi
347             WY0[w[d,n]] <- WY0[w[d,n]] - 1
348             # calcolo la prob. di avere yV[d,n]=0 e yV[d,n]=1
349             p_Yv0 <- (bV[2] + N-1 - Yv1) * (betaV[w[d,n]] + WY0[w[d,n]]) / (
    betaV_sum + sum(WY0))
350             p_Yv1 <- (bV[1] + Yv1) * (betaV[w[d,n]] + WY1ZX[w[d,n],zV[d,n]]) / (
    betaV_sum + sum(WY1ZX[,zV[d,n]]))
351             # aggiorno yV[d,n]
352             yV[d,n] <- sample(c(0,1), size=1, prob=c(p_Yv0,p_Yv1)/sum(p_Yv0,p_
    Yv1))
353             # aggiorno i conteggi
354             if (yV[d,n]==1) {
355               WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + 1
356               Yv1 <- Yv1 + 1
357               WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
358             } else {
359               WY0[w[d,n]] <- WY0[w[d,n]] + 1
360             }
361           } else {
362             # rimuovo yV[d,n] dai conteggi
363             WY0[w[d,n]] <- WY0[w[d,n]] - 1
364             # calcolo la prob. di avere yV[d,n]=0 e yV[d,n]=1
365             p_Yv0 <- (bV[2] + N-1 - Yv1) * (betaV[w[d,n]] + WY0[w[d,n]]) / (
    betaV_sum + sum(WY0))
366             p_Yv1 <- (bV[1] + Yv1) * (betaV[w[d,n]] + WY1ZX[w[d,n],zstar[d]]) /
    (betaV_sum + sum(WY1ZX[,zstar[d]]))
367             # aggiorno yV[d,n]
368             yV[d,n] <- sample(c(0,1), size=1, prob=c(p_Yv0,p_Yv1)/sum(p_Yv0,p_
    Yv1))
369             # aggiorno i conteggi
370             if (yV[d,n]==1) {
371               WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
372               Yv1 <- Yv1 + 1
373               WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
374             } else {
375               WY0[w[d,n]] <- WY0[w[d,n]] + 1
376             }
377           }
378         }
379         # FINE AGGIORNAMENTO yV[d,n]
380         # ------------------------------------------------------------------ #
381         # AGGIORNAMENTO zV[d,n]
382         if (yV[d,n]==1) {
383           if (x[d]==1) {
384             # rimuovo zV[d,n] dai conteggi
385             WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] - 1
```

```
386             Z[d,zV[d,n]] <- Z[d,zV[d,n]] - 1
387             WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] - 1
388             # calcolo la prob. di avere zV[d,n]=t per t=1,...,T
389             p_zV <- (alpha0+lambda[d,]*alpha + Z[d,]) * (betaV[w[d,n]] + WY1ZX[w
    [d,n],]) / (betaV_sum + apply(WY1ZX,2,sum))
390             # aggiorno zV[d,n]
391             zV[d,n] <- sample.int(TOPICS, size=1, replace=T, prob=p_zV/sum(p_zV)
    )
392             # aggiorno i conteggi
393             WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + 1
394             Z[d,zV[d,n]] <- Z[d,zV[d,n]] + 1
395             WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
396           } else {
397             # rimuovo zV[d,n] dai conteggi
398             WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] - 1
399             Z[d,zV[d,n]] <- Z[d,zV[d,n]] - 1
400             WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] - 1
401             # calcolo la prob. di avere zV[d,n]=t per t=1,...,T
402             p_zV <- (alpha0+lambda[d,]*alpha + Z[d,])
403             # aggiorno zV[d,n]
404             zV[d,n] <- sample.int(TOPICS, size=1, replace=T, prob=p_zV/sum(p_zV)
    )
405             # aggiorno i conteggi
406             WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
407             Z[d,zV[d,n]] <- Z[d,zV[d,n]] + 1
408             WY1Z[[d]][w[d,n],zV[d,n]] <- WY1Z[[d]][w[d,n],zV[d,n]] + 1
409           }
410         } else {
411           # rimuovo zV[d,n] dai conteggi
412           Z[d,zV[d,n]] <- Z[d,zV[d,n]] - 1
413           # calcolo la prob. di avere zV[d,n]=t per t=1,...,T
414           p_zV <- (alpha0+lambda[d,]*alpha + Z[d,])
415           # aggiorno zV[d,n]
416           zV[d,n] <- sample.int(TOPICS, size=1, replace=T, prob=p_zV/sum(p_zV))
417           # aggiorno i conteggi
418           Z[d,zV[d,n]] <- Z[d,zV[d,n]] + 1
419         }
420         # FINE AGGIORNAMENTO zV[d,n]
421         # ------------------------------------------------------------------ #
422       }
423       for (l in Ld_range) {
424         # ------------------------------------------------------------------ #
425         # AGGIORNAMENTO yH[d,l]
426         if (yH[d,l]==1) {
427           if (x[d]==1) {
428             # rimuovo yH[d,l] dai conteggi
429             HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] - 1
430             Yh1 <- Yh1 - 1
431             HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] - 1
432             # calcolo la prob. di avere yH[d,l]=0 e yH[d,l]=1
```

```
433             p_Yh0 <- (bH[2] + L-1 - Yh1) * (betaH[h[d,l]] + HY0[h[d,l]]) / (
    betaH_sum + sum(HY0))
434             p_Yh1 <- (bH[1] + Yh1) * (betaH[h[d,l]] + HY1ZX[h[d,l],zH[d,l]]) / (
    betaH_sum + sum(HY1ZX[,zH[d,l]]))
435             # aggiorno yV[d,n]
436             yH[d,l] <- sample(c(0,1), size=1, prob=c(p_Yh0,p_Yh1)/sum(p_Yh0,p_
    Yh1))
437             # aggiorno i conteggi
438             if (yH[d,l]==1) {
439               HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] + 1
440               Yh1 <- Yh1 + 1
441               HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
442             } else {
443               HY0[h[d,l]] <- HY0[h[d,l]] + 1
444             }
445           } else {
446             # rimuovo yH[d,l] dai conteggi
447             HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] - 1
448             Yh1 <- Yh1 - 1
449             HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] - 1
450             # calcolo la prob. di avere yH[d,l]=0 e yH[d,l]=1
451             p_Yh0 <- (bH[2] + L-1 - Yh1) * (betaH[h[d,l]] + HY0[h[d,l]]) / (
    betaH_sum + sum(HY0))
452             p_Yh1 <- (bH[1] + Yh1) * (betaH[h[d,l]] + HY1ZX[h[d,l],zstar[d]]) /
    (betaH_sum + sum(HY1ZX[,zstar[d]]))
453             # aggiorno yV[d,n]
454             yH[d,l] <- sample(c(0,1), size=1, prob=c(p_Yh0,p_Yh1)/sum(p_Yh0,p_
    Yh1))
455             # aggiorno i conteggi
456             if (yH[d,l]==1) {
457               HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
458               Yh1 <- Yh1 + 1
459               HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
460             } else {
461               HY0[h[d,l]] <- HY0[h[d,l]] + 1
462             }
463           }
464         } else {
465           if (x[d]==1) {
466             # rimuovo yH[d,l] dai conteggi
467             HY0[h[d,l]] <- HY0[h[d,l]] - 1
468             # calcolo la prob. di avere yH[d,l]=0 e yH[d,l]=1
469             p_Yh0 <- (bH[2] + L-1 - Yh1) * (betaH[h[d,l]] + HY0[h[d,l]]) / (
    betaH_sum + sum(HY0))
470             p_Yh1 <- (bH[1] + Yh1) * (betaH[h[d,l]] + HY1ZX[h[d,l],zH[d,l]]) / (
    betaH_sum + sum(HY1ZX[,zH[d,l]]))
471             # aggiorno yV[d,n]
472             yH[d,l] <- sample(c(0,1), size=1, prob=c(p_Yh0,p_Yh1)/sum(p_Yh0,p_
    Yh1))
473             # aggiorno i conteggi
```

```
474             if (yH[d,l]==1) {
475               HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] + 1
476               Yh1 <- Yh1 + 1
477               HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
478             } else {
479               HY0[h[d,l]] <- HY0[h[d,l]] + 1
480             }
481           } else {
482             # rimuovo yH[d,l] dai conteggi
483             HY0[h[d,l]] <- HY0[h[d,l]] - 1
484             # calcolo la prob. di avere yH[d,l]=0 e yH[d,l]=1
485             p_Yh0 <- (bH[2] + L-1 - Yh1) * (betaH[h[d,l]] + HY0[h[d,l]]) / (
    betaH_sum + sum(HY0))
486             p_Yh1 <- (bH[1] + Yh1) * (betaH[h[d,l]] + HY1ZX[h[d,l],zstar[d]]) /
    (betaH_sum + sum(HY1ZX[,zstar[d]]))
487             # aggiorno yV[d,n]
488             yH[d,l] <- sample(c(0,1), size=1, prob=c(p_Yh0,p_Yh1)/sum(p_Yh0,p_
    Yh1))
489             # aggiorno i conteggi
490             if (yH[d,l]==1) {
491               HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
492               Yh1 <- Yh1 + 1
493               HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
494             } else {
495               HY0[h[d,l]] <- HY0[h[d,l]] + 1
496             }
497           }
498         }
499         # FINE AGGIORNAMENTO yH[d,l]
500         # ------------------------------------------------------------------ #
501         # AGGIORNAMENTO zH[d,l]
502         if (yH[d,l]==1) {
503           if (x[d]==1) {
504             # rimuovo zH[d,l] dai conteggi
505             HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] - 1
506             Z[d,zH[d,l]] <- Z[d,zH[d,l]] - 1
507             HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] - 1
508             # calcolo la prob. di avere zH[d,l]=t per t=1,...,T
509             p_zH <- (alpha0+lambda[d,]*alpha + Z[d,]) * (betaH[h[d,l]] + HY1ZX[h
    [d,l],]) / (betaH_sum + apply(HY1ZX,2,sum))
510             # aggiorno zH[d,l]
511             zH[d,l] <- sample.int(TOPICS, size=1, replace=T, prob=p_zH/sum(p_zH)
    )
512             # aggiorno i conteggi
513             HY1ZX[h[d,l],zH[d,l]] <- HY1ZX[h[d,l],zH[d,l]] + 1
514             Z[d,zH[d,l]] <- Z[d,zH[d,l]] + 1
515             HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
516           } else {
517             # rimuovo zH[d,l] dai conteggi
518             HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] - 1
```

```
519             Z[d,zH[d,l]] <- Z[d,zH[d,l]] - 1
520             HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] - 1
521             # calcolo la prob. di avere zH[d,l]=t per t=1,...,T
522             p_zH <- (alpha0+lambda[d,]*alpha + Z[d,])
523             # aggiorno zH[d,l]
524             zH[d,l] <- sample.int(TOPICS, size=1, replace=T, prob=p_zH/sum(p_zH)
    )
525             # aggiorno i conteggi
526             HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
527             Z[d,zH[d,l]] <- Z[d,zH[d,l]] + 1
528             HY1Z[[d]][h[d,l],zH[d,l]] <- HY1Z[[d]][h[d,l],zH[d,l]] + 1
529           }
530         } else {
531           # rimuovo zH[d,l] dai conteggi
532           Z[d,zH[d,l]] <- Z[d,zH[d,l]] - 1
533           # calcolo la prob. di avere zV[d,n]=t per t=1,...,T
534           p_zH <- (alpha0+lambda[d,]*alpha + Z[d,])
535           # aggiorno zV[d,n]
536           zH[d,l] <- sample.int(TOPICS, size=1, replace=T, prob=p_zH/sum(p_zH))
537           # aggiorno i conteggi
538           Z[d,zH[d,l]] <- Z[d,zH[d,l]] + 1
539         }
540         # FINE AGGIORNAMENTO zH[d,l]
541         # ------------------------------------------------------------------- #
542       }
543       # FINE AGGIORNAMENTO DOCUMENTO
544       # ------------------------------------------------------------------- #
545     }
546     # salvo m-mo stato in una serie di file
547     saveRDS(x,      file=file.path(result_folder, "x", paste("x_", str_pad(m,3,
     pad="0"),".RDS", sep="")))
548     saveRDS(zstar,  file=file.path(result_folder, "zstar", paste("zstar_", str_
     pad(m,3,pad="0"),".RDS", sep="")))
549     saveRDS(lambda, file=file.path(result_folder, "lambda", paste("lambda_", str
     _pad(m,3,pad="0"),".RDS", sep="")))
550     saveRDS(yV,     file=file.path(result_folder, "yV", paste("yV_", str_pad(m
     ,3,pad="0"),".RDS", sep="")))
551     saveRDS(zV,     file=file.path(result_folder, "zV", paste("zV_", str_pad(m
     ,3,pad="0"),".RDS", sep="")))
552     saveRDS(yH,     file=file.path(result_folder, "yH", paste("yH_", str_pad(m
     ,3,pad="0"),".RDS", sep="")))
553     saveRDS(zH,     file=file.path(result_folder, "zH", paste("zH_", str_pad(m
     ,3,pad="0"),".RDS", sep="")))
554     # FINE ITERAZIONE CGS
555   }
556   # FINE COLLAPSED GIBBS SAMPLING
557   # ------------------------------------------------------------------------- #
558   cat("\n", as.character(Sys.time()), " FINE", sep="")
559   # FINE FUNZIONE
560   # ------------------------------------------------------------------------- #
```

```
561 }
562
563 # ------------------------------------------------------------------------ #
```

CGS.R

C.3 Collapsed Gibbs Sampler (LDA)

```
# ------------------------------------------------------------------------ #

CGS_LDA <- function(w, alpha, betaV, iterations=100, seed=28, result_folder) {
  # ------------------------------------------------------------------------ #
  # Argomenti della funzione:
  #          w : matrice D x Nmax   | n-ma parola del d-mo documento
  #      alpha : vettore TOPICS x 1 | parametro Dirichlet sul simplesso dei topic
  #      betaV : vettore V x 1      | parametro Dirichlet sul simplesso delle parole
  # iterations : intero             | numero di stati della catena da campionare
  #       seed : intero             | seme per rendere i risultati replicabili
  # ------------------------------------------------------------------------ #
  cat("\n", as.character(Sys.time()), " Operazioni preliminari.", sep="")
  # Importo librerie
  require(Matrix)
  require(stringr)
  # Fisso il seme
  set.seed(seed)
  # Creo cartella in cui salvare gli stati della catena
  result_folder <- file.path(getwd(), "results", result_folder)
  if(!dir.exists(result_folder)) {
    dir.create(result_folder)
    dir.create(file.path(result_folder, "zV"))
  } else {
    cat("\nLa cartella '",result_folder,"' esiste gia': selezionare un altro valore per la variabile 'result_folder'.\n", sep="")
    return(NULL)
  }
  # Definisco alcune quantita' utili
  TOPICS <- length(alpha)
  D <- nrow(w)
  V <- length(betaV)
  N <- sum(w>0)
  betaV_sum <- sum(betaV)
  # ------------------------------------------------------------------------ #
  # GENERAZIONE MATRICI DI CONTEGGI
  cat("\n", as.character(Sys.time()), " Creazione delle matrici di conteggi.", sep="")
  # Creo le matrici dei conteggi
  WY1ZX <- matrix(0, nrow=V, ncol=TOPICS)
  Z <- matrix(0, nrow=D, ncol=TOPICS)
  # FINE GENERAZIONE MATRICI DI CONTEGGI
  # ------------------------------------------------------------------------ #
  # GENERAZIONE STATO INIZIALE
  cat("\n", as.character(Sys.time()), " Generazione dello stato iniziale: ", sep="")
```

```
43  # Genero le var. dello stato iniziale e in contemporanea aggiorno i conteggi
44  zV <- matrix(0, nrow=D, ncol=ncol(w))
45  # ------------------------------------------------------------------------ #
46  for (d in 1:D) {
47    if (d %in% round(quantile(1:D,(1:10)/10))) cat("-")
48    Nd_range <- if(sum(w[d,]>0)>0) 1:sum(w[d,]>0) else c()  # indici delle
       parole nel documento
49    for (n in Nd_range) {
50      # estraggo valori
51      zV[d,n] <- sample.int(TOPICS, size=1, replace=T, prob=alpha/sum(alpha))
52      # aggiorno i conteggi
53      Z[d,zV[d,n]] <- Z[d,zV[d,n]] + 1
54      WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + 1
55    }
56  }
57  # FINE GENERAZIONE STATO INIZIALE
58  # ------------------------------------------------------------------------ #
59  # COLLAPSED GIBBS SAMPLING (CGS)
60  cat("\n", as.character(Sys.time()), " Collapsed Gibbs Sampler (", iterations,
       " iterazioni)", sep="")
61  for (m in 1:iterations) {
62    # ------------------------------------------------------------------------ #
63    # ITERAZIONE CGS
64    cat("\n", as.character(Sys.time()), "  - iterazione ", str_pad(m, 3, pad = "
        "), ": ", sep="")
65    for (d in 1:D) {
66      # ------------------------------------------------------------------------ #
67      # AGGIORNAMENTO DOCUMENTO
68      if (d %in% round(quantile(1:D,(1:20)/20))) cat("-")
69      # definisco alcune quantita' utili
70      Nd_range <- if(sum(w[d,]>0)>0) 1:sum(w[d,]>0) else c()  # indici delle
       parole nel documento
71      # ------------------------------------------------------------------------ #
72      for (n in Nd_range) {
73        # ------------------------------------------------------------------------ #
74        # AGGIORNAMENTO zV[d,n]
75        # rimuovo zV[d,n] dai conteggi
76        WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] - 1
77        Z[d,zV[d,n]] <- Z[d,zV[d,n]] - 1
78        # calcolo la prob. di avere zV[d,n]=t per t=1,...,T
79        p_zV <- (alpha + Z[d,]) * (betaV[w[d,n]] + WY1ZX[w[d,n],]) / (betaV_sum
       + apply(WY1ZX,2,sum))
80        # aggiorno zV[d,n]
81        zV[d,n] <- sample.int(TOPICS, size=1, replace=T, prob=p_zV/sum(p_zV))
82        # aggiorno i conteggi
83        WY1ZX[w[d,n],zV[d,n]] <- WY1ZX[w[d,n],zV[d,n]] + 1
84        Z[d,zV[d,n]] <- Z[d,zV[d,n]] + 1
85        # FINE AGGIORNAMENTO zV[d,n]
86        # ------------------------------------------------------------------------ #
87      }
```

```
88        # FINE AGGIORNAMENTO DOCUMENTO
89        # ------------------------------------------------------------------------ #
90      }
91      # salvo m-mo stato in una serie di file
92      saveRDS(zV, file=file.path(result_folder, "zV", paste("zV_", str_pad(m,3,pad
        ="0"),".RDS", sep="")))
93      # FINE ITERAZIONE CGS
94    }
95    # FINE COLLAPSED GIBBS SAMPLING
96    # ------------------------------------------------------------------------ #
97    cat("\n", as.character(Sys.time()), " FINE", sep="")
98    # FINE FUNZIONE
99    # ------------------------------------------------------------------------ #
100 }
101
102 # ------------------------------------------------------------------------ #
```

CGS_LDA.R

C.4 Collapsed Gibbs Sampler (Twitter-LDA)

```
# ------------------------------------------------------------------------- #

CGS_TwitterLDA <- function(w, doc_users, alphastar, betaV, bV,
                           iterations=300, seed=28, result_folder) {
  # ------------------------------------------------------------------------- #
  # Argomenti della funzione:
  #          w : matrice D x Nmax   | n-ma parola del d-mo documento
  #  doc_users : vettore D x 1      | autore del d-mo topic
  #  alphastar : vettore TOPICS x 1 | parametro Dirichlet sul simplesso dei topic
  #      betaV : vettore V x 1      | parametro Dirichlet sul simplesso delle parole
  #         bV : vettore 2x1        | parametro Beta
  # iterations : intero             | numero di stati della catena da campionare
  #       seed : intero             | seme per rendere i risultati replicabili
  # ------------------------------------------------------------------------- #
  cat("\n", as.character(Sys.time()), " Operazioni preliminari.", sep="")
  # Importo librerie
  require(Matrix)
  require(stringr)
  # Fisso il seme
  set.seed(seed)
  # Creo cartella in cui salvare gli stati della catena
  result_folder <- file.path(getwd(), "results", result_folder)
  if(!dir.exists(result_folder)) {
    dir.create(result_folder)
    dir.create(file.path(result_folder, "zstar"))
    dir.create(file.path(result_folder, "yV"))
  } else {
    cat("\nLa cartella '",result_folder,"' esiste gia': selezionare un altro valore per la variabile 'result_folder'.\n", sep="")
    return(NULL)
  }
  # Definisco alcune quantita' utili
  TOPICS <- length(alphastar)
  U <- length(unique(doc_users))
  D <- nrow(w)
  V <- length(betaV)
  N <- sum(w>0)
  betaV_sum <- sum(betaV)
  # ------------------------------------------------------------------------- #
  # GENERAZIONE MATRICI DI CONTEGGI
  cat("\n", as.character(Sys.time()), " Creazione delle matrici di conteggi.", sep="")
  # Creo le matrici dei conteggi
  WY1ZX <- matrix(0, nrow=V, ncol=TOPICS)
  Zstar <- matrix(0, nrow=U, ncol=TOPICS)
```

```
44  Yv1 <- 0
45  WY0 <- rep(0, V)
46  WY1 <- matrix(0, nrow=D, ncol=V)
47  # FINE GENERAZIONE MATRICI DI CONTEGGI
48  # ------------------------------------------------------------------------ #
49  # GENERAZIONE STATO INIZIALE
50  cat("\n", as.character(Sys.time()), " Generazione dello stato iniziale: ", sep
      ="")
51  # Genero le var. dello stato iniziale e in contemporanea aggiorno i conteggi
52  zstar <- rep(0, D)
53  yV <- matrix(0, nrow=D, ncol=ncol(w))
54  # ------------------------------------------------------------------------ #
55  for (d in 1:D) {
56    if (d %in% round(quantile(1:D,(1:10)/10))) cat("-")
57    u <- doc_users[d]            # autore del documento
58    Nd_range <-  1:sum(w[d,]>0)  # indici delle parole nel documento
59    # estraggo valori
60    zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=alphastar/sum(
       alphastar))
61    # aggiorno i conteggi
62    Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
63    for (n in Nd_range) {
64      # estraggo valori
65      yV[d,n] <- rbinom(1, size=1, p=bV[1]/sum(bV))
66      # aggiorno i conteggi
67      if (yV[d,n]==1) {
68        WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
69        Yv1 <- Yv1 + 1
70        WY1[d,w[d,n]] <- WY1[d,w[d,n]] + 1
71      } else {
72        WY0[w[d,n]] <- WY0[w[d,n]] + 1
73      }
74    }
75  }
76  # FINE GENERAZIONE STATO INIZIALE
77  # ------------------------------------------------------------------------ #
78  # COLLAPSED GIBBS SAMPLING (CGS)
79  cat("\n", as.character(Sys.time()), " Collapsed Gibbs Sampler (", iterations,
       " iterazioni)", sep="")
80  for (m in 1:iterations) {
81    # ------------------------------------------------------------------------ #
82    # ITERAZIONE CGS
83    cat("\n", as.character(Sys.time()), "  - iterazione ", str_pad(m, 3, pad = "
        "), ": ", sep="")
84    for (d in 1:D) {
85      # ------------------------------------------------------------------------ #
86      # AGGIORNAMENTO DOCUMENTO
87      if (d %in% round(quantile(1:D,(1:20)/20))) cat("-")
88      # definisco alcune quantita' utili
89      u <- doc_users[d]                      # autore del documento
```

```
90      Nd_range <-  1:sum(w[d,]>0)            # indici delle parole nel documento
91      unique_words <- unique(w[d,Nd_range])  # parole distinte nel documento
92      # ------------------------------------------------------------------------ #
93      # AGGIORNAMENTO zstar[d]
94      # rimuovo zstar[d] dai conteggi
95      Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] - 1
96      for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] - yV[
    d,n]
97      # calcolo la prob. di avere zstar[d]=t
98      p_zstar <- alphastar + Zstar[u,]
99      for (t in 1:TOPICS) {
100       qmax  <- sum(WY1[d,])
101       if (qmax>0) {
102         p_num <- c(); p_den <- betaV_sum + sum(WY1ZX[,t]) + 0:(qmax-1)
103         for (w_dn in unique_words) {
104           if (WY1[d,w_dn]>0) p_num <- c(p_num, betaV[w_dn] + WY1ZX[w_dn,t] +
    0:(WY1[d,w_dn]-1))
105         }
106         p_zstar[t] <- p_zstar[t] * prod(p_num/p_den)
107       }
108     }
109     # aggiorno zstar[d]
110     zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=p_zstar/sum(p_zstar
    ))
111     # aggiorno i conteggi
112     Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
113     for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + yV[
    d,n]
114     # FINE AGGIORNAMENTO zstar[d]
115     # ------------------------------------------------------------------------ #
116     for (n in Nd_range) {
117       # ---------------------------------------------------------------------- #
118       # AGGIORNAMENTO yV[d,n]
119       if (yV[d,n]==1) {
120         # rimuovo yV[d,n] dai conteggi
121         WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] - 1
122         Yv1 <- Yv1 - 1
123         WY1[d,w[d,n]] <- WY1[d,w[d,n]] - 1
124         # calcolo la prob. di avere yV[d,n]=0 e yV[d,n]=1
125         p_Yv0 <- (bV[2] + N-1 - Yv1) * (betaV[w[d,n]] + WY0[w[d,n]]) / (betaV_
    sum + sum(WY0))
126         p_Yv1 <- (bV[1] + Yv1) * (betaV[w[d,n]] + WY1ZX[w[d,n],zstar[d]]) / (
    betaV_sum + sum(WY1ZX[,zstar[d]]))
127         # aggiorno yV[d,n]
128         yV[d,n] <- sample(c(0,1), size=1, prob=c(p_Yv0,p_Yv1)/sum(p_Yv0,p_Yv1)
    )
129         # aggiorno i conteggi
130         if (yV[d,n]==1) {
131           WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
132           Yv1 <- Yv1 + 1
```

```
133             WY1[d,w[d,n]] <- WY1[d,w[d,n]] + 1
134           } else {
135             WY0[w[d,n]] <- WY0[w[d,n]] + 1
136           }
137         } else {
138           # rimuovo yV[d,n] dai conteggi
139           WY0[w[d,n]] <- WY0[w[d,n]] - 1
140           # calcolo la prob. di avere yV[d,n]=0 e yV[d,n]=1
141           p_Yv0 <- (bV[2] + N-1 - Yv1) * (betaV[w[d,n]] + WY0[w[d,n]]) / (betaV_
      sum + sum(WY0))
142           p_Yv1 <- (bV[1] + Yv1) * (betaV[w[d,n]] + WY1ZX[w[d,n],zstar[d]]) / (
      betaV_sum + sum(WY1ZX[,zstar[d]]))
143           # aggiorno yV[d,n]
144           yV[d,n] <- sample(c(0,1), size=1, prob=c(p_Yv0,p_Yv1)/sum(p_Yv0,p_Yv1)
      )
145           # aggiorno i conteggi
146           if (yV[d,n]==1) {
147             WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
148             Yv1 <- Yv1 + 1
149             WY1[d,w[d,n]] <- WY1[d,w[d,n]] + 1
150           } else {
151             WY0[w[d,n]] <- WY0[w[d,n]] + 1
152           }
153         }
154         # FINE AGGIORNAMENTO yV[d,n]
155         # ------------------------------------------------------------------- #
156       }
157       # FINE AGGIORNAMENTO DOCUMENTO
158       # ------------------------------------------------------------------- #
159     }
160     # salvo m-mo stato in una serie di file
161     saveRDS(zstar,  file=file.path(result_folder, "zstar", paste("zstar_", str_
      pad(m,3,pad="0"),".RDS", sep="")))
162     saveRDS(yV,     file=file.path(result_folder, "yV", paste("yV_", str_pad(m
      ,3,pad="0"),".RDS", sep="")))
163     # FINE ITERAZIONE CGS
164   }
165   # FINE COLLAPSED GIBBS SAMPLING
166   # ------------------------------------------------------------------------ #
167   cat("\n", as.character(Sys.time()), " FINE", sep="")
168   # FINE FUNZIONE
169   # ------------------------------------------------------------------------ #
170 }
171
172 # ------------------------------------------------------------------------ #
```

CGS_TwitterLDA.R

C.5 Collapsed Gibbs Sampler (Hashtag-LDA)

```
# ------------------------------------------------------------------------- #

CGS_HashtagLDA <- function(w, h, doc_users, alphastar, betaV, betaH, bH,
                           iterations=300, seed=28, result_folder) {
  # ------------------------------------------------------------------------- #
  # Argomenti della funzione:
  #          w : matrice D x Nmax   | n-ma parola del d-mo documento
  #          h : matrice D x Lmax   | l-mo hashtag del d-mo documento
  #  doc_users : vettore D x 1      | autore del d-mo topic
  #  alphastar : vettore TOPICS x 1 | parametro Dirichlet sul simplesso dei topic
  #      betaV : vettore V x 1      | parametro Dirichlet sul simplesso delle parole
  #      betaH : vettore H x 1      | parametro Dirichlet sul simplesso degli hashtag
  #         bH : vettore 2x1        | parametro Beta
  # iterations : intero             | numero di stati della catena da campionare
  #       seed : intero             | seme per rendere i risultati replicabili
  # ------------------------------------------------------------------------- #
  cat("\n", as.character(Sys.time()), " Operazioni preliminari.", sep="")
  # Importo librerie
  require(Matrix)
  require(stringr)
  # Fisso il seme
  set.seed(seed)
  # Creo cartella in cui salvare gli stati della catena
  result_folder <- file.path(getwd(), "results", result_folder)
  if(!dir.exists(result_folder)) {
    dir.create(result_folder)
    dir.create(file.path(result_folder, "zstar"))
    dir.create(file.path(result_folder, "yH"))
  } else {
    cat("\nLa cartella '",result_folder,"' esiste gia': selezionare un altro valore per la variabile 'result_folder'.\n", sep="")
    return(NULL)
  }
  # Definisco alcune quantita' utili
  TOPICS <- length(alphastar)
  U <- length(unique(doc_users))
  D <- nrow(w)
  V <- length(betaV)
  H <- length(betaH)
  N <- sum(w>0)
  L <- sum(h>0)
  betaV_sum <- sum(betaV)
  betaH_sum <- sum(betaH)
  # ------------------------------------------------------------------------- #
```

```
44  # GENERAZIONE MATRICI DI CONTEGGI
45  cat("\n", as.character(Sys.time()), " Creazione delle matrici di conteggi.",
       sep="")
46  # Creo le matrici dei conteggi
47  WY1ZX <- matrix(0, nrow=V, ncol=TOPICS)
48  HY1ZX <- matrix(0, nrow=H, ncol=TOPICS)
49  Zstar <- matrix(0, nrow=U, ncol=TOPICS)
50  Yh1 <- 0
51  HY0 <- rep(0, H)
52  WY1 <- matrix(0, nrow=D, ncol=V)
53  HY1 <- matrix(0, nrow=D, ncol=H)
54  # FINE GENERAZIONE MATRICI DI CONTEGGI
55  # ------------------------------------------------------------------------ #
56  # GENERAZIONE STATO INIZIALE
57  cat("\n", as.character(Sys.time()), " Generazione dello stato iniziale: ", sep
       ="")
58  # Genero le var. dello stato iniziale e in contemporanea aggiorno i conteggi
59  zstar <- rep(0, D)
60  yH <- matrix(0, nrow=D, ncol=ncol(h))
61  # ------------------------------------------------------------------------ #
62  for (d in 1:D) {
63    if (d %in% round(quantile(1:D,(1:10)/10))) cat("-")
64    u <- doc_users[d]   # autore del documento
65    Nd_range <- if(sum(w[d,]>0)>0) 1:sum(w[d,]>0) else c()  # indici delle
       parole nel documento
66    Ld_range <- if(sum(h[d,]>0)>0) 1:sum(h[d,]>0) else c()  # indici degli
       hashtag nel documento
67    # estraggo valori
68    zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=alphastar/sum(
       alphastar))
69    # aggiorno i conteggi
70    Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
71    for (n in Nd_range) {
72      # aggiorno i conteggi
73      WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
74      WY1[d,w[d,n]] <- WY1[d,w[d,n]] + 1
75    }
76    for (l in Ld_range) {
77      # estraggo valori
78      yH[d,l] <- rbinom(1, size=1, p=bH[1]/sum(bH))
79      # aggiorno i conteggi
80      if (yH[d,l]==1) {
81        HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
82        Yh1 <- Yh1 + 1
83        HY1[d,h[d,l]] <- HY1[d,h[d,l]] + 1
84      } else {
85        HY0[h[d,l]] <- HY0[h[d,l]] + 1
86      }
87    }
88  }
```

```
89  # FINE GENERAZIONE STATO INIZIALE
90  # ------------------------------------------------------------------------ #
91  # COLLAPSED GIBBS SAMPLING (CGS)
92  cat("\n", as.character(Sys.time()), " Collapsed Gibbs Sampler (", iterations,
      " iterazioni)", sep="")
93  for (m in 1:iterations) {
94    # ------------------------------------------------------------------------ #
95    # ITERAZIONE CGS
96    cat("\n", as.character(Sys.time()), "  - iterazione ", str_pad(m, 3, pad = "
        "), ": ", sep="")
97    for (d in 1:D) {
98      # ------------------------------------------------------------------------ #
99      # AGGIORNAMENTO DOCUMENTO
100     if (d %in% round(quantile(1:D,(1:20)/20))) cat("-")
101     # definisco alcune quantita' utili
102     u <- doc_users[d]   # autore del documento
103     Nd <- sum(w[d,]>0)
104     if(Nd>0) {
105       Nd_range <-  1:Nd                        # indici delle parole nel
      documento
106       unique_words <- unique(w[d,Nd_range])    # parole distinte nel documento
107     } else {
108       Nd_range <- unique_words <- c()
109     }
110     Ld <- sum(h[d,]>0)
111     if(Ld>0) {
112       Ld_range <- 1:Ld                         # indici degli hashtag nel
      documento
113       unique_hashtags <- unique(h[d,Ld_range]) # hashtag distinti nel
      documento
114     } else {
115       Ld_range <- unique_hashtags <- c()
116     }
117     # ------------------------------------------------------------------------ #
118     # AGGIORNAMENTO zstar[d]
119     # rimuovo zstar[d] dai conteggi
120     Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] - 1
121     for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] - 1
122     for (l in Ld_range) HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] - yH[
      d,l]
123     # calcolo la prob. di avere zstar[d]=t
124     p_zstar <- alphastar + Zstar[u,]
125     for (t in 1:TOPICS) {
126       qmax  <- sum(WY1[d,])
127       if (qmax>0) {
128         p_num <- c(); p_den <- betaV_sum + sum(WY1ZX[,t]) + 0:(qmax-1)
129         for (w_dn in unique_words) {
130           if (WY1[d,w_dn]>0) p_num <- c(p_num, betaV[w_dn] + WY1ZX[w_dn,t] +
      0:(WY1[d,w_dn]-1))
131         }
```

```
132           p_zstar[t] <- p_zstar[t] * prod(p_num/p_den)
133         }
134         qmax  <- sum(HY1[d,])
135         if (qmax>0) {
136           p_num <- c(); p_den <- betaH_sum + sum(HY1ZX[,t]) + 0:(qmax-1)
137           for (h_dl in unique_hashtags) {
138             if (HY1[d,h_dl]>0) p_num <- c(p_num, betaH[h_dl] + HY1ZX[h_dl,t] +
    0:(HY1[d,h_dl]-1))
139           }
140           p_zstar[t] <- p_zstar[t] * prod(p_num/p_den)
141         }
142       }
143       # aggiorno zstar[d]
144       zstar[d] <- sample.int(TOPICS, size=1, replace=T, prob=p_zstar/sum(p_zstar
    ))
145       # aggiorno i conteggi
146       Zstar[u,zstar[d]] <- Zstar[u,zstar[d]] + 1
147       for (n in Nd_range) WY1ZX[w[d,n],zstar[d]] <- WY1ZX[w[d,n],zstar[d]] + 1
148       for (l in Ld_range) HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + yH[
    d,l]
149       # FINE AGGIORNAMENTO zstar[d]
150       # ------------------------------------------------------------------------ #
151       for (l in Ld_range) {
152         # ------------------------------------------------------------------------ #
153         # AGGIORNAMENTO yH[d,l]
154         if (yH[d,l]==1) {
155           # rimuovo yH[d,l] dai conteggi
156           HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] - 1
157           Yh1 <- Yh1 - 1
158           HY1[d,h[d,l]] <- HY1[d,h[d,l]] - 1
159           # calcolo la prob. di avere yH[d,l]=0 e yH[d,l]=1
160           p_Yh0 <- (bH[2] + L-1 - Yh1) * (betaH[h[d,l]] + HY0[h[d,l]]) / (betaH_
    sum + sum(HY0))
161           p_Yh1 <- (bH[1] + Yh1) * (betaH[h[d,l]] + HY1ZX[h[d,l],zstar[d]]) / (
    betaH_sum + sum(HY1ZX[,zstar[d]]))
162           # aggiorno yH[d,n]
163           yH[d,l] <- sample(c(0,1), size=1, prob=c(p_Yh0,p_Yh1))
164           # aggiorno i conteggi
165           if (yH[d,l]==1) {
166             HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
167             Yh1 <- Yh1 + 1
168             HY1[d,h[d,l]] <- HY1[d,h[d,l]] + 1
169           } else {
170             HY0[h[d,l]] <- HY0[h[d,l]] + 1
171           }
172         } else {
173           # rimuovo yH[d,l] dai conteggi
174           HY0[h[d,l]] <- HY0[h[d,l]] - 1
175           # calcolo la prob. di avere yH[d,l]=0 e yH[d,l]=1
```

```
176           p_Yh0 <- (bH[2] + L-1 - Yh1) * (betaH[h[d,l]] + HY0[h[d,l]]) / (betaH_
      sum + sum(HY0))
177           p_Yh1 <- (bH[1] + Yh1) * (betaH[h[d,l]] + HY1ZX[h[d,l],zstar[d]]) / (
      betaH_sum + sum(HY1ZX[,zstar[d]]))
178           # aggiorno yH[d,n]
179           yH[d,l] <- sample(c(0,1), size=1, prob=c(p_Yh0,p_Yh1)/sum(p_Yh0,p_Yh1)
      )
180           # aggiorno i conteggi
181           if (yH[d,l]==1) {
182             HY1ZX[h[d,l],zstar[d]] <- HY1ZX[h[d,l],zstar[d]] + 1
183             Yh1 <- Yh1 + 1
184             HY1[d,h[d,l]] <- HY1[d,h[d,l]] + 1
185           } else {
186             HY0[h[d,l]] <- HY0[h[d,l]] + 1
187           }
188         }
189         # FINE AGGIORNAMENTO yH[d,l]
190         # ------------------------------------------------------------------------ #
191       }
192       # FINE AGGIORNAMENTO DOCUMENTO
193       # ------------------------------------------------------------------------ #
194     }
195     # salvo m-mo stato in una serie di file
196     saveRDS(zstar,  file=file.path(result_folder, "zstar", paste("zstar_", str_
      pad(m,3,pad="0"),".RDS", sep="")))
197     saveRDS(yH,     file=file.path(result_folder, "yH", paste("yH_", str_pad(m
      ,3,pad="0"),".RDS", sep="")))
198     # FINE ITERAZIONE CGS
199   }
200   # FINE COLLAPSED GIBBS SAMPLING
201   # ------------------------------------------------------------------------------ #
202   cat("\n", as.character(Sys.time()), " FINE", sep="")
203   # FINE FUNZIONE
204   # ------------------------------------------------------------------------------ #
205 }
206
207 # ------------------------------------------------------------------------------ #
```

CGS_HashtagLDA.R

C.6 Metriche di Topic Coherence

```
# ------------------------------------------------------------------------ #

TC_ALL <- function(phi, top=20, freq_vector) {
  # ------------------------------------------------------------------------ #
  #         phi : matrice TOPICS x V | distribuzione sui termini dei T topic
  #         top : intero             | numero di top word da considerare
  # freq_vector : matrice V x D      | indica se il v-mo termine e' nel d-mo doc
  # ------------------------------------------------------------------------ #
  # definisco la funzione NZ
  PMI <- function(f1, f2) {
    # ------------------------------------------------------------------------ #
    #  f1 : doc. in cui compare w1
    #  f2 : doc. in cui compare w2
    # f12 : doc. in cui compaiono sia w1 sia w2
    # ------------------------------------------------------------------------ #
    f12 <- sum(f1 * f2)
    if (f12==0) return(0)
    log( f12 * length(f1) / sum(f1) / sum(f2))
  }
  LCP <- function(f1, f2, smoothing=1) {
    # ------------------------------------------------------------------------ #
    #  f1 : doc. in cui compare w1
    #  f2 : doc. in cui compare w2
    # f12 : doc. in cui compaiono sia w1 sia w2
    # ------------------------------------------------------------------------ #
    f12 <- sum(f1 * f2)
    log((f12+smoothing)/sum(f2))
  }
  NZ <- function(f1, f2) {
    # ------------------------------------------------------------------------ #
    #  f1 : doc. in cui compare w1
    #  f2 : doc. in cui compare w2
    # f12 : doc. in cui compaiono sia w1 sia w2
    # ------------------------------------------------------------------------ #
    f12 <- f1 * f2
    sum(f12) == 0
  }
  # definisco quantita' utili
  TOPICS <- nrow(phi)
  top <- min(ncol(phi), top)
  out_PMI <- out_NPMI <- out_LCP <- out_NZ <- rep(0, TOPICS)
  # calcolo TC-PMI, TC-LCP e TC-NZ
  for (t in 1:TOPICS) {
    top_words <- order(phi[t,], decreasing=T)[1:top]
    for (i in 2:top) {
      for (j in 1:(i-1)) {
        out_PMI[t] <- out_PMI[t] + PMI(freq_vector[top_words[i],],
```

```
                                    freq_vector[top_words[j],])
        out_LCP[t] <- out_LCP[t] + LCP(freq_vector[top_words[i],],
                                    freq_vector[top_words[j],])
        out_NZ[t] <- out_NZ[t] + NZ(freq_vector[top_words[i],],
                                  freq_vector[top_words[j],])
      }
    }
  }
  list("PMI" = out_PMI * 2 / (top^2-top),
       "LCP" = out_LCP * 2 / (top^2-top),
       "NZ" = out_NZ * 2 / (top^2-top))
}

# ------------------------------------------------------------------------- #
```

topic_coherence.R

Appendice D

Codice Python

D.1 Librerie

Le librerie di *Python* utilizzate in questa tesi sono:

- `emoji` Kim e Wurster, n.d.
- `pandas` pandas development team, 2020 McKinney, 2010
- `tweepy` Roesslein, n.d.

D.2 Download dei Tweet

```
# ------------------------------------------------------------------------ #

import tweepy
import time
import pandas as pd
from datetime import datetime, timedelta
from emoji import demojize

# ------------------------------------------------------------------------ #

bearer_token = ...
client = tweepy.Client(bearer_token=bearer_token)

# ------------------------------------------------------------------------ #

keywords = ("sars-cov-2", "covid19", "sars-cov2", "pandemia",
            "covid", '"corona virus"', "coronavirus")

# query
query = "(" + " OR ".join(keywords) + ") lang:it -is:retweet"

```

```
# ------------------------------------------------------------------------ #

# data e orario
start_time = "2022-01-30T21:00:00Z"
end_time = "2022-02-05T21:00:00Z"

# ------------------------------------------------------------------------ #

tws = tweepy.Paginator(client.search_recent_tweets, query=query,
                       start_time=start_time, end_time=end_time,
                       tweet_fields=["author_id","conversation_id",
                                     "created_at","in_reply_to_user_id",
                                     "public_metrics","lang"],
                       max_results=100).flatten(limit=20000)

tweet0 = []
for tw in tws:
    tweet0.append([str(tw.id),
                   str(tw.author_id),
                   str(tw.conversation_id),
                   tw.text, tw.created_at,
                   str(tw.in_reply_to_user_id),
                   tw.public_metrics["retweet_count"],
                   tw.public_metrics["reply_count"],
                   tw.public_metrics["like_count"],
                   tw.public_metrics["quote_count"],
                   tw.lang])

# ------------------------------------------------------------------------ #

colnames = ["tweet_id", "author_id", "conversation_id", "text", "created_at",
            "in_reply_to_user_id", "retweet_count", "reply_count",
            "like_count", "quote_count", "lang"]

tweet0 = pd.DataFrame(tweet0, columns=colnames)
tweet0.to_csv("data/csv/tweet0_original.csv", sep="\t",
              index=False, encoding="utf-8-sig", mode="w")

# ------------------------------------------------------------------------ #

c_id_list = list(set(tweet0["conversation_id"]))
n = 10  # numero di conversazioni da considerare nella stessa query
conversation_id_blocks = [c_id_list[i:i + n] for i in range(0, len(c_id_list), n
    )]

tweets = []
i = 0
while i < len(conversation_id_blocks):
    query = "(conversation_id:" + " OR conversation_id:".join(conversation_id_
    blocks[i]) + ") lang:it"
```

```
70      try:
71          tws = tweepy.Paginator(client.search_recent_tweets, query=query,
72                                 start_time=start_time, end_time=end_time,
73                                 tweet_fields=["author_id","conversation_id",
74                                               "created_at","in_reply_to_user_id",
75                                               "public_metrics","lang"],
76                                 max_results=100).flatten(limit=100000)
77          for tw in tws:
78              tweets.append([str(tw.id), str(tw.author_id), str(tw.conversation_id
        ), tw.text, tw.created_at,
79                             str(tw.in_reply_to_user_id),
80                             tw.public_metrics["retweet_count"],tw.public_metrics[
        "reply_count"],
81                             tw.public_metrics["like_count"],tw.public_metrics["
        quote_count"],
82                             tw.lang])
83          i += 1
84          print(i, ":", len(tweets))
85      except:
86          print("PAUSETTA DI 15 MINUTI.", end=" ")
87          time.sleep(300)
88          print("-10", end=" ")
89          time.sleep(300)
90          print("-5")
91          time.sleep(300)
92
93  # Converto la lista di liste in un dataframe
94  tweets = pd.DataFrame(tweets, columns=colnames)
95  tweets.to_csv("data/csv/tweets_original.csv",
96                sep="\t", index=False, encoding="utf-8-sig", mode="w")
97
98  # ------------------------------------------------------------------------ #
99
100 tweet0tweets = pd.concat([tweet0,tweets])
101 print("Dimensione originale:      ", tweet0tweets.shape)
102
103 tweet0tweets = tweet0tweets.drop_duplicates(subset=["tweet_id"])
104 print("Dimensione senza duplicati:", tweet0tweets.shape)
105
106 # ------------------------------------------------------------------------ #
107
108 tweet0tweets.to_csv("data/csv/tweet0tweets_original.csv",
109                     sep="\t", index=False, encoding="utf-8-sig", mode="w")
110
111 # ------------------------------------------------------------------------ #
112
113 tweet0tweets["text"] = tweet0tweets["text"].apply(lambda x: demojize(x, language
        ="en", delimiters=(" emote_"," ")))
114
115 # ------------------------------------------------------------------------ #
```

```
116
117 tweet0tweets.to_csv("data/csv/tweet0tweets.csv",
118                      sep="\t", index=False, encoding="utf-8-sig", mode="w")
119
120 # ------------------------------------------------------------------------ #
```

Bibliografia

AlSumait, L., Barbará, D., Gentle, J. & Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models. *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, 67–82.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics* [R package version 2.3]. https://CRAN.R-project.org/package=gridExtra

Bates, D. & Maechler, M. (2021). *Matrix: Sparse and Dense Matrix Classes and Methods* [R package version 1.3-3]. https://CRAN.R-project.org/package=Matrix

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag.

Blei, D. M. (2012). Probabilistic Topic Models. *Commun. ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Carin, L. & Dunson, D. (2010). Probabilistic Topic Models. *IEEE Signal Processing Magazine*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Blei, D. M. & Lafferty, J. D. (2006). Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyd-Graber, J., Mimno, D. & Newman, D. (2014). Handbook of Mixed Membership Models and Their Applications. In E. M. Airoldi, D. M. Blei, E. A. Erosheva & S. E. Fienberg (Cur.). Chapman; Hall.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 288–296.

Cohen, S. (2019). Bayesian Analysis in Natural Language Processing, Second Edition. *Synthesis Lectures on Human Language Technologies*, *12*(1), 1–343. https://doi.org/10.2200/S00905ED2V01Y201903HLT041

Consortium, T. U. (n.d.). *Full Emoji List, v13.1*. Recuperato febbraio 26, 2022, da http://www.unicode.org/emoji/charts/full-emoji-list.html

Friedman, D. (2019). *topicdoc: Topic-Specific Diagnostics for LDA and CTM Topic Models* [R package version 0.1.0]. https://CRAN.R-project.org/package=topicdoc

Gao, J. & Johnson, M. (2008). A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers. *EMNLP 2008*, 344–352.

Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Griffiths, T. L., Steyvers, M., Blei, D. M. & Tenenbaum, J. (2005). Integrating Topics and Syntax. In L. Saul, Y. Weiss & L. Bottou (Cur.), *Advances in Neural Information Processing Systems*. MIT Press.

Grün, B. & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. https://doi.org/10.1145/312624.312649

Hong, L. & Davison, B. D. (2010). Empirical Study of Topic Modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics*, 80–88. https://doi.org/10.1145/1964858.1964870

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimedia Tools Appl.*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kaplan, A. M. & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, *54*(2), 105–113. https://doi.org/https://doi.org/10.1016/j.bushor.2010.09.004

Kim, T. & Wurster, K. (n.d.). *Emoji for Python*. Recuperato marzo 3, 2022, da https://pypi.org/project/emoji

Lang, D. & Chien, G.-t. (2018). *wordcloud2: Create Word Cloud by 'htmlwidget'* [R package version 0.2.1]. https://CRAN.R-project.org/package=wordcloud2

Lin, T., Tian, W., Mei, Q. & Cheng, H. (2014). The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text. *Proceedings of the 23rd International Conference on World Wide Web*, 539–550. https://doi.org/10.1145/2566486.2567980

Ma, Z., Dou, W., Wang, X. & Akella, S. (2013). Tag-Latent Dirichlet Allocation: Understanding Hashtags and Their Relationships. *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*, 260–267. https://doi.org/10.1109/WI-IAT.2013.38

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Cur.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. (2013). Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 889–892. https://doi.org/10.1145/2484028.2484166

Meschiari, S. (2022). *latex2exp: Use LaTeX Expressions in Plots* [R package version 0.9.3]. https://CRAN.R-project.org/package=latex2exp

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Newman, D., Lau, J. H., Grieser, K. & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108.

Nigam, K., McCallum, A. K., Thrun, S. & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*(2), 103–134.

Pace, L. & Salvan, A. (2001). *Introduzione alla Statistica - II. Inferenza, Verosimiglianza, Modelli*. Cedam.

pandas development team, T. (2020). *pandas-dev/pandas: Pandas 1.0.3* (Ver. v1.0.3). Zenodo. https://doi.org/10.5281/zenodo.3715232

Resnik, P. & Hardisty, E. (2010). *Gibbs sampling for the uninitiated* (rapp. tecn.). University of Maryland College Park Institution for Advanced Computer Studies.

Roesslein, J. (n.d.). *Tweepy: Twitter for Python!* Recuperato febbraio 25, 2022, da https://github.com/tweepy/tweepy

Rosen-Zvi, M., Griffiths, T. L., Steyvers, M. & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487–494.

Silge, J. & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3). https://doi.org/10.21105/joss.00037

Solari, D., Sciandra, A. & Finos, L. (2019). TextWiller: Collection of functions for text mining, specially devoted to the Italian language. *Journal of Open Source Software*, *4*(41), 1256. https://doi.org/10.21105/joss.01256

Srivastava, A. N. & Sahami, M. (2009). Topic Models. In D. M. Blei & J. D. Lafferty (Cur.), *Text mining: Classification, clustering, and applications*. Chapman; Hall.

Steyvers, M. & Griffiths, T. L. (2007). Latent Semantic Analysis: A Road to Meaning. In T. Landauer, S. D. McNamara & W. Kintsch (Cur.). Laurence Erlbaum.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Tsai, F. S. (2011). A Tag-Topic Model for Blog Mining. *Expert Syst. Appl.*, *38*(5), 5330–5335. https://doi.org/10.1016/j.eswa.2010.10.025

Twitter. (n.d.-a). *Informazioni sulle API di Twitter*. Recuperato febbraio 3, 2022, da https://help.twitter.com/en/rules-and-policies/twitter-api

Twitter. (n.d.-b). *Search Tweets - How to build a query*. Recuperato febbraio 4, 2022, da https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

Twitter. (n.d.-c). *Search Tweets introduction*. Recuperato febbraio 4, 2022, da https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

Wang, X. & McCallum, A. (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 424–433. https://doi.org/10.1145/1150402.1150450

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations* [R package version 1.4.0]. https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L. & Müller, K. (2021). *dplyr: A Grammar of Data Manipulation* [R package version 1.0.7]. https://CRAN.R-project.org/package=dplyr

Zhao, F., Zhu, Y., Jin, H. & Yang, L. T. (2016). A Personalized Hashtag Recommendation Approach Using LDA-Based Topic Model in Microblog Environment. *Future Gener. Comput. Syst.*, *65*(100), 196–206. https://doi.org/10.1016/j.future.2015.10.012

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. & Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee & V. Mudoch (Cur.), *Advances in Information Retrieval* (pp. 338–349). Springer Berlin Heidelberg.