Riassunti degli articoli

Giovanni Toto

13 novembre 2021

Indice

1 Modelli generativi & Topic Models

Un *modello generativo* specifica una procedura statistica secondo cui i documenti possono essere generati; le leggi statistiche che definiscono il modello si possono basare sull'utilizzo di variabili latenti. In particolare, quando si stima un modello generativo, si vuole trovare la miglior combinazione di variabili latenti che possa spiegare i dati osservati –le parole nei documenti–, assumendo che il modello abbia effettivamente generato i dati (Steyvers e T. Griffiths 2007). In pratica, viene fissato un processo generativo ragionevole, poi si inverte il processo attraverso procedure statistiche e si fa inferenza sull'insieme di topic responsabili della generazione della collezione di documenti.

Topic Models –Misture di unigrammi, *pLSI*, *LDA*, *...*– sono *modelli generativi probabilistici* che si basano sull'idea che ogni documento di una collezione è modellato come una mistura di topic e ogni topic è caratterizzato da una distribuzione di probabilità sulle parole di un vocabolario noto e fissato. In particolare, il processo generativo di un documento è il seguente: si sceglie una distribuzione sui topic, per ogni parola del documento è scelto un topic in base alla distribuzione scelta e infine ogni parola è estratta a partire dalla distribuzione del topic assegnato alla parola (Steyvers e T. Griffiths 2007).

In questo caso, le parole che compongono i documenti sono le variabili osservate, mentre tutto ciò che è legato ai topic –proporzione dei topic di ogni documento, distribuzione delle parole per ogni topic, topic assegnato a ogni parola– va a costituire la struttura latente. Il problema principale dei *topic model* è quindi fare inferenza sui componenti della struttura latente avendo a disposizione solo i documenti testuali (D. M. Blei 2012).

1.1 Obiettivi & Applicazioni

L'obiettivo della *LDA* –e più in generale dei *topic model*– è ottenere una breve descrizione di ogni documento di una collezione che permetta preservare le relazioni statistiche essenziali. Queste risultano utili per effettuare, ad esempio, classificazione, collaborative filtering, valutare la similarità tra documenti o tra parole, valutare la rilevanza di documenti rispetto a una query, ... (D. M. Blei, Ng e Jordan 2003, Steyvers e T. Griffiths 2007).

È importante sottolineare che i *topic model* sono metodi non supervisionati, permettono quindi di processare efficientemente grandi moli di dati in maniera completamente automatica ed ad una velocità impossibile per annotatori umani.

Si rimanda all'articolo di Boyd-Graber, Hu e Mimno 2017 per un'introduzione ai *topic model* e una trattazione dettagliata delle loro principali applicazioni.

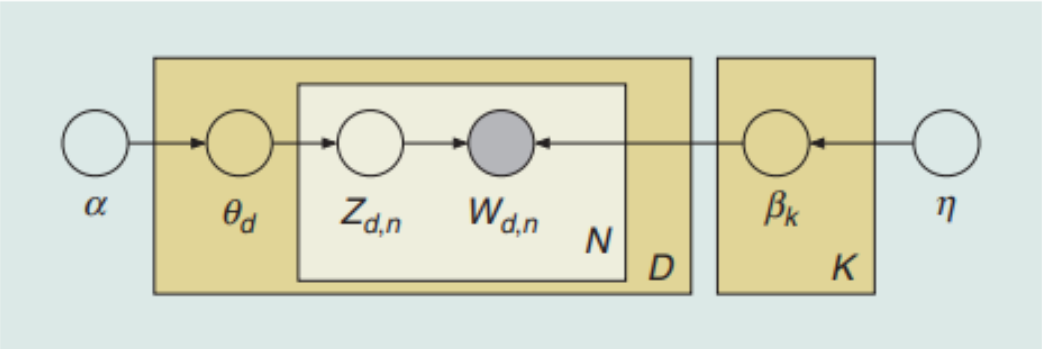
2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (*LDA*) è un modello bayesiano gerarchico su tre livelli; in particolare, è composto da una gerarchia di modelli mistura in cui ogni documento è modellato come una mistura finita di topic in cui i pesi –sotto definiti *proporzioni dei topic*– sono estratti una volta

sola per ogni documento ma i componenti della mistura, ovvero i topic, sono condivisi da tutti i documenti della collezione (D. Blei, Carin e Dunson 2010).

La notazione e l'immagine[1] sono tratti da D. Blei, Carin e Dunson 2010. Siano K il numero di topic, V il numero di parole nel vocabolario, D il numero di documenti e N il numero di parole in un documento, il processo generativo di *smoothed LDA*[2] è il seguente:

1. Campiono K distribuzioni sulle V parole, una per ogni topic, da una distribuzione di Dirichlet simmetrica, $\beta_k \sim Dir_V(\eta)$, $k = 1, \ldots, K$.
2. Per ogni documento d, campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta_d \sim Dir_K(\alpha)$, $d = 1, \ldots, D$.
3. Per ogni parola n in ogni documento d:
   a. Estraggo un topic dalle proporzioni dei topic, $z_{d,n}|\theta_d \sim Mult(\theta_d)$.
   b. Estraggo una parola dal topic corrispondente, $w_{d,n}|z_{d,n}, \beta_{1:K} \sim Mult(\beta_{z_{d,n}})$.



[FIG2] The graphical model representation of LDA [15].

Il processo generativo definisce un distribuzione congiunta di variabili osservate e variabili latenti, tuttavia per fare inferenza si è interessati alla distribuzione delle variabili latenti date quelle osservate:

$$\Pr(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}|w_{1:D,1:N})$$

Effettuare inferenza esatta non è possibile, quindi si ricorrono ad approssimazioni ottenute tramite *MCMC*, *Gibbs Sampling*, *Variational Inference*, ...

2.1 Interpretazione della Distribuzione a Posteriori

Una volta calcolata la distribuzione a posteriori si ottengono le matrici $\beta_{1:K} = (\beta_1 \ldots \beta_K)^\top \in \mathbb{R}^{K \times V}$ e $\theta_{1:D} = (\theta_1 \ldots \theta_D)^\top \in \mathbb{R}^{D \times K}$; in particolare:

- La k-ma riga di $\beta_{1:K}$, β_k, è la distribuzione di probabilità sulle V parole del k-mo topic; basta andare a vedere le parole a cui corrispondono probabilità più alte per valutare di cosa il topic tratta effettivamente.
- La v-colonna di $\beta_{1:K}$ contiene la probabilità di osservare la v-ma parola nei K topic; permette di valutare in quali topic una parola è più utilizzata e in quali meno.

[1] Come interpretare l'immagine è spiegato bene in Steyvers e T. Griffiths 2007.
[2] Nella versione standard si assegna una distribuzione a priori solo a θ, nella versione *smoothed* anche a β.

- La d-ma riga di $\theta_{1:D}$, θ_d, è la proporzione dei K topic nel d-mo documento; permette di valutare di quali topic il d-mo documento tratta e in che proporzione.
- La k-ma colonna di $\theta_{1:D}$ contiene le proporzioni del k-mo topic nei D documenti; permette di valutare quanto il k-mo topic è trattato all'interno della collezione.

2.2 Assunzioni di base

In D. M. Blei 2012 vengono analizzate tre assunzioni di base della *LDA*:

- Numero di topic T noto e fissato a priori.
- Assunzione *bag-of-words* o assunzione di scambiabilità[3] delle parole all'interno di un documento, secondo cui non è rilevante l'ordine della parole all'interno di un documento ma solo la loro frequenza.
- Assunzione di scambiabilità dei documenti all'interno della collezione, secondo cui non è rilevante considerare l'ordinamento temporale dei testi.

3 Estensioni della LDA

Uno dei principali vantaggi della *LDA* è la sua modularità che permette di formulare abbastanza facilmente estensioni del modello originale; più nello specifico molte di queste sono ottenute rilassando una o più delle tre assunzioni precedenti. Ciò è ragionevole poiché in alcuni casi possono essere considerate troppo stringenti. Di seguito, prima si riportano modelli che rilassano le assunzioni precedenti, successivamente si introducono vari modelli che utilizzano metadati, poi un modello che gestisce la sparsità nelle proporzioni dei topic e nelle distribuzioni delle parole in testi brevi, infine un modello che considera le correlazioni tra topic.

Per ulteriori modelli, si rimanda all'articolo di Jelodar et al. 2019 in cui vengono trattate le principali estensioni della *LDA* pubblicate tra il 2003 e il 2016.

3.1 Numero di topic

La scelta del numero di topic T può influenzare fortemente l'interpretabilità dei risultati; in particolare, 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 di T:

- Utilizzare una conoscenza a priori del dataset.
- Stimare modelli con diversi T e poi selezionarne uno attraverso un criterio fissato; solitamente si usa la *perplessità* o la più alta probabilità a posteriori.

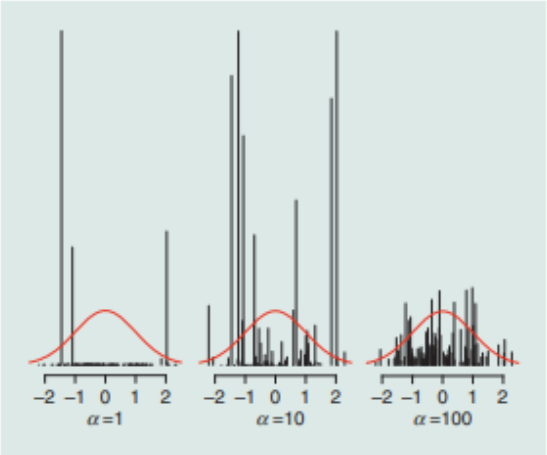
[3]Una sequenza finita di variabili aleatorie $\{z_1, \ldots, z_N\}$ si dice scambiabile se la distribuzione congiunta è invariante a permutazioni; una sequenza infinita si dice infinitamente scambiabile se ogni sua sotto-sequenza finita è scambiabile.

- Metodi non parametrici in cui il modello seleziona autonomamente il numero appropriato di topic, includendo la scelta all'interno della procedura di inferenza della distribuzione a posteriori.

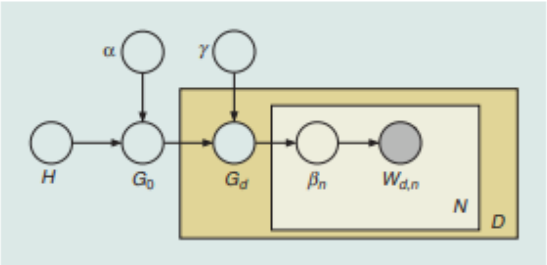
Teh et al. 2006 propongono un metodo bayesiano non parametrico, detto *Hierarchical Dirichlet Process*[4] (*HDP*), che permette di ottenere un *topic model* con un numero di topic infinito a priori; in particolare, il numero dei topic diventa parte della distribuzione a posteriori dalla struttura latente. Per costruire il nuovo modello, le proporzioni dei topic θ_d 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*.

1. Campiono una distribuzione di base sulle proporzioni dei topic, $G_0|\gamma \sim DP(\gamma, H)$ dove H è una distribuzione di Dirichlet simmetrica sul simplesso delle parole.
2. Per ogni documento $d = 1, \ldots, D$:
   (a) Campiono le proporzioni dei topic, $G_d|\alpha, G_0 \sim DP(\alpha, G_0)$.
   (b) Per ogni parola $n = 1, \ldots, N$:
      i. Estraggo un topic dalle proporzioni dei topic, $\beta_{dn}|G_d \sim G_d$.
      ii. Estraggo una parola dal topic corrispondente, $w_{dn}|\beta_{dn} \sim Mult(\beta_{dn})$.

I parametri γ e α dei *DP* determinano quanto è probabile osservare nuove osservazioni, dette *atomi*: con valori bassi si avranno pochi atomi con quasi la totalità della probabilità mentre con valori alti si avrà che la distribuzione del *DP* tenderà a coincidere con quella di base[5], H e G_0 rispettivamente (D. Blei, Carin e Dunson 2010). G_0 può essere una qualsiasi distribuzione continua sui topic, tuttavia in questo caso è fondamentale che la distribuzione di base sia un *DP* poiché ciò permette di avere topic –che corrispondono agli atomi del processo– condivisi tra tutti i documenti della collezione; considerando invece una qualsiasi altra distribuzione, si avrebbero solo le parole condivise tra i documenti e non i topic. Una proprietà molto attrattiva di questo nuovo metodo –che rende ragionevole applicarlo a collezioni in continua evoluzione– è che nuovi documenti possono generare nuovi topic mai visti in precedenza e questa generazione emerge come conseguenza naturale del modello probabilistico (D. Blei, Carin e Dunson 2010).



[FIG3] Three draws from a DP with standard normal base distribution. Draws from the DP are discrete; as α increases, the resulting random distribution looks more like the base distribution.



[FIG4] The graphical model representation of the HDP topic model [54].

[4]Si veda l'Appendice A per un'analisi più approfondita.

[5]L'idea di fondo è che si ha una distribuzione continua ma si vuole avere probabilità non nulla di estrarre più volte lo stesso valore; di norma, ciò non è possibile poiché $\Pr(X = x) = 0$ se X è una variabile casuale continua.

3.2 Scambiabilità tra parole

Questa assunzione è irrealistica ma è ragionevole per ragioni computazionali e inferenziali, e se l'unico obiettivo è rivelare la struttura semantica latente dei testi attraverso i topic.

3.2.1 Composite model (2005)

Thomas Griffiths et al. 2005 sviluppano un *topic model* che considera sia le dipendenze sintattiche tra parole vicine all'interno di una frase sia le dipendenze semantiche tra parole anche lontane all'interno di un documento; in particolare, utilizzano un *modello composito* (*composite model*) che permette di gestire il fatto che tutte le parole presentano dipendenze sintattiche ma solo alcune anche dipendenze semantiche. A ogni parola è assegnata una classe e l'appartenenza a una determinata classe dipende dalla classe della parola precedente attraverso un *Hidden Markov Model* (*HMM*); alle parole assegnate a una particolare classe, detta *classe semantica*, viene attribuito un topic attraverso un *topic model*. Il d-mo documento è quindi generato della seguente procedura:

1. Campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta^{(d)} \sim Dir_K(\alpha)$.
   a. Estraggo un topic dalle proporzioni dei topic, $z_i \sim \theta^{(d)}$.
   b. Estraggo una classe dalla proporzione di transizione da c_{i-1} a c_i, $c_i \sim \pi^{(c_i-1)}$.
   c. Se $c_i = 1$, allora estraggo w_i da $\phi^{(z_i)}$; altrimenti estraggo w_i da $\phi^{(c_i)}$.

Per effettuare l'inferenza bayesiana si utilizza *MCMC*, assumendo le seguenti distribuzioni a priori per i parametri: $\theta^{(d)} \sim Dir_K(\alpha)$, $\phi^{(z)} \sim Dir_V(\beta)$, $\phi^{(c)} \sim Dir_V(\delta)$ e $Dir_C(\gamma)$ per le righe delle matrici di transizione del *HMM*.

3.2.2 Bigram Topic Model (2006)

Wallach 2006 sviluppa il *Bigram Topic Model* in cui i topic generano le parole condizionatamente alla parola precedente; in particolare, combina un *n-gram model*, un modello che genera una parola a partire dalle n precedenti[6], e un *topic model* che genera una parola sulla base di topic latenti determinati solo a partire dalle correlazioni tra parole, ignorando quindi l'ordinamento. Essenzialmente si considera un *LDA* in cui la generazione delle parole non dipende più esclusivamente dal topic assegnato ma anche dal contesto, ovvero dalla parola precedente; si ha quindi che ogni topic è caratterizzato da V distribuzioni, una per ogni possibile parola precedente:

$$\beta_{1:VK} = (\beta_{1,1} \dots \beta_{V,1} \dots \beta_{1,K} \dots \beta_{V,K})^\top \in \mathbb{R}^{VK \times V}$$

Si passa quindi da una matrice con $K(V-1)$ a una con $VK(V-1)$ parametri liberi[7]. Per le righe $\beta_{j,k}$ della matrice è possibile considerare diverse distribuzioni a priori in base a come si vuole condividere l'informazione a priori tra topic e contesti: stessa distribuzione a priori per tutte le possibili combinazioni di topic e contesto, stessa distribuzione a priori per $\beta_{1,k}, \beta_{2,k}, \dots, \beta_{V,k}$ (stesso topic k) o stessa distribuzione a priori per $\beta_{j,1}, \beta_{j,2}, \dots, \beta_{j,K}$ (stesso contesto j).

[6] Wallach 2006 considera un *n-gram model* con $n = 2$, ovvero un *bigram model*, per costruire il *Bigram Topic Model*; il modello può essere esteso considerando un qualsiasi $n \in \mathbb{N}$.

[7] Si ha $V - 1$ poiché l'ultimo valore di ogni riga è ottenibile come complemento a 1 degli altri $V - 1$.

3.3 Scambiabilità tra documenti

L'assunzione secondo cui l'ordine dei documenti non è rilevante può essere ragionevole per collezioni che raccolgono documenti circa tutti dello 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.

3.3.1 dDTM (2006) & cDTM (2008)

D. M. Blei e Lafferty 2006 propongono il *Dynamic Topic Model* (*DTM* o *dDTM*) in cui si rilassa l'assunzione di scambiabilità dei documenti all'interno della collezione, assumendo 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. Per modellare l'evoluzione della distribuzione di probabilità sulle V parole del k-mo topic all'epoca t si assume che $\beta_{t,k}$ abbia distribuzione logit-normale in modo da poter usare un modello state-space con rumore gaussiano[8]: $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I_V)$;

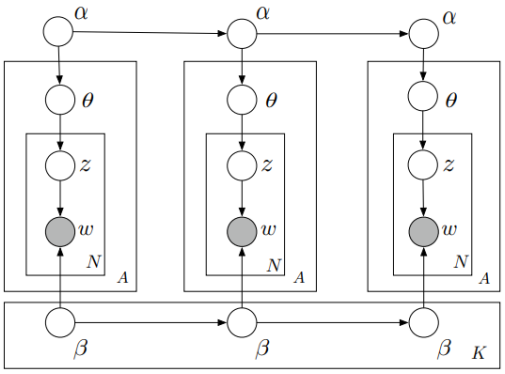


Figure 1. Graphical representation of a dynamic topic model (for three time slices). Each topic's natural parameters $\beta_{t,k}$ evolve over time, together with the mean parameters α_t of the logistic normal distribution for the topic proportions.

Si ha quindi una sequenza di *topic model* –uno per ogni epoca– collegati attraverso i modelli state-space sopra definiti; il processo generativo per l'epoca $t = 1$ coincide con l'LDA, mentre per $t > 1$ è il seguente:

1. Campiono K distribuzioni sulle V parole introducendo rumore a quelle dell'epoca precedente, $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I_V)$, $k = 1, \ldots, K$.
2. Per ogni documento d:
   a. Campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta_d \sim Dir_K(\alpha)$.
   b. Per ogni parola n nel documento d:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{t,d,n}|\theta_d \sim Mult(\theta_d)$.
      ii. Estraggo una parola dal topic corrispondente $w_{t,d,n}|z_{t,d,n}, \beta_{t,1:K} \sim Mult(\pi(\beta_{t,z_{t,d,n}}))$.

dove $\pi(\cdot) = \text{softmax}(\cdot)$ rende il vettore $\beta_{t,k}$ utilizzabile come il vettore dei parametri di una distribuzione multinomiale; il w-mo componente di $\pi(\beta_{t,k})$ è dato da:

$$\pi(\beta_{t,k})_w = \frac{\exp(\beta_{t,k,w})}{\sum_w \exp(\beta_{t,k,w})}$$

Citanto D. Blei, Carin e Dunson 2010, il modello permette di:

[8] L'idea di base è aggiornare il vettore precedente con del rumore gaussiano e poi trasformarlo nei parametri di una distribuzione multinomiale attraverso la funzione softmax; in particolare, il vettore ottenuto ha componenti in $(0, 1)$ che sommano a 1.

- investigare come cambia un topic k col passare del tempo osservando le parole con probabilità più alta ad ogni epoca, ovvero osservando $\beta_{1,k}, \beta_{2,k}, \ldots$;
- investigare come l'importanza di una parola w cambia col passare del tempo all'interno dello stesso topic k, osservando $\beta_{1,k,w}, \beta_{2,k,w}, \ldots$;
- esaminare documenti associati a un determinato topic in epoche differenti.

Le limitazioni principali del *dDTM* sono dovute alla discretizzazione del tempo e in particolare al numero di intervalli considerati, detti *epoche*: se sono troppo pochi, l'assunzione di scambiabilità al loro interno non risulta più ragionevole; se sono troppi, si ha un'eccessiva complessità nell'*inferenza variazionale* (C. Wang, D. Blei e Heckerman 2008).

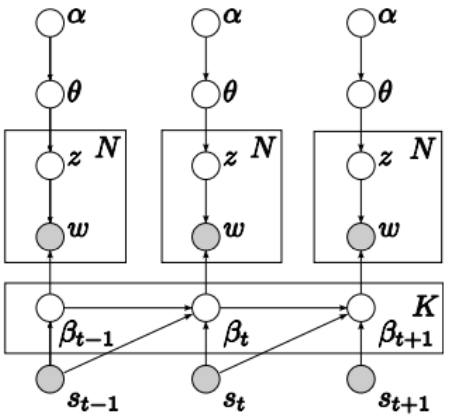


Figure 1: Graphical model representation of the cDTM. The evolution of the topic parameters β_t is governed by Brownian motion. The variable s_t is the observed time stamp of document d_t.

C. Wang, D. Blei e Heckerman 2008 propongono un'estensione del *dDTM* detta *continuous Dynamic Topic Model* (*CDTM*) che risolve i problemi sopra citati considerando il tempo come continuo. Il modello state-space discreto viene sostituito con la sua generalizzazione continua, ovvero un *moto Browniano*; in particolare, sia assume che il componente relativo alla parola w nella distribuzione delle V parole del k-mo topic sia

$$\beta_{0,k,w} \sim N(m, v_0)$$
$$\beta_{j,k,w}|\beta_{i,k,w}, s \sim N(\beta_{i,k,w}, v\Delta_{s_j,s_i})$$

dove Δ_{s_j,s_i} è il tempo trascorso –lag– tra due timestamp relativi rispettivamente agli indici temporali i e j, $j > i > 0$; si ha quindi che la varianza aumenta linearmente con il lag. Il processo generativo è:

1. Per ogni topic $k = 1, \ldots, K$:
   a. Campiono la distribuzione iniziale delle V parole del k-mo topic, $\beta_{0,k} \sim N(m, v_0 I_V)$.
2. Per ogni documento d_t al tempo s_t:
   a. Per ogni topic $k = 1, \ldots, K$:
      i. Aggiorno la distribuzione delle V parole del k-mo topic, $\beta_{t,k}|\beta_{t-1,k}, s \sim N(\beta_{i,k}, v\Delta_{s_t,s_{t-1}} I_v)$
   b. Campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta_t \sim Dir_K(\alpha)$.
   c. Per ogni parola w:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{t,n}|\theta_t \sim Mult(\theta_t)$.
      ii. Estraggo una parola dal topic corrispondente $w_{t,n}|z_{t,n}, \beta_{t,1:K} \sim Mult(\pi(\beta_{t,z_{t,n}}))$.

A differenza del *dDTM*, il *cDTM* aggiorna la distribuzione delle V parole $\beta_{t,1:K}$ per ogni timestamp s_t osservato: se due documenti presentano lo stesso timestamp, la distribuzione non varia poiché

la varianza del moto Browniano si annulla. Si noti che è possibile ottenere un *dDTM* a partire da un *cDTM* quando la varianza del modello state-space tende a 0, ovvero se $\sigma^2 \to 0$ nella notazione di D. M. Blei e Lafferty 2006. Per l'inferenza dei parametri si riprende il *variational Kalman filtering algorithm* del *dDTM* e lo si adatta al caso continuo sostituendo il modello state-space con il *moto Browniano*, inoltre si ricorre a una procedura sparsa[9] in modo da limitare il numero di parametri da stimare e quindi ridurre la complessità del problema di inferenza.

3.3.2 TOT (2006) & npTOT (2013)

X. Wang e McCallum 2006 propongono *Topic Over Time* (*TOT*), un modello generativo che modella congiuntamente la collocazione nel tempo dei documenti –rappresentata da un timestamp– e le co-occorrenze delle parole; a differenza di altri lavori dello stesso periodo, il modello non discretizza[10] il tempo né considera processi Markoviani per modellare l'evoluzione dei parametri; inoltre vengono catturati i cambiamenti nelle co-occorenze tra parole nei topic, non i cambiamenti nella distribuzione delle parole di ogni topic al variare del tempo.

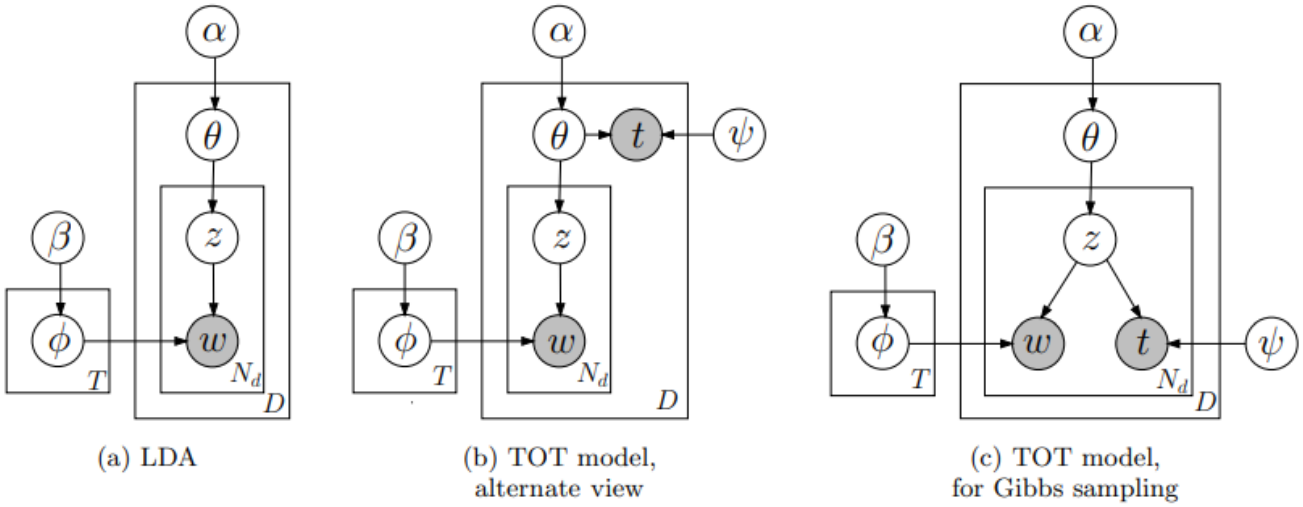


Figure 1: Three topic models: LDA and two perspectives on TOT

Nella figura tratta da X. Wang e McCallum 2006, sono riportati l'*LDA* e i due modelli proposti nell'articolo: (b) genera un unico timestamp t_d per ogni documento mentre (c) genera un timestamp t_{di} per ogni parola. (c) è il processo generativo usato nel *Gibbs sampler* per la stima della distribuzione a posteriori ed è:

1. Per ogni topic $z = 1, \ldots, T$:
   (a) Campiono la distribuzione sulle V parole del topic da una distribuzione di Dirichlet simmetrica, $\phi_z \sim Dir_V(\beta)$.
   (b) Scelgo un insieme di parametri ψ_z per parametrizzare la distribuzione Beta.

[9] L'idea di base è che se non si osserva la parola w al tempo t, allora la vera distribuzione a posteriori di $\beta_{t,w}$ è determinata solo dalle altre parole osservate al tempo t e quindi non è necessario rappresentare i parametri variazionali $\hat{\beta}_{t,w}$ per le parole w non osservate (C. Wang, D. Blei e Heckerman 2008).

[10] Per evitare la discretizzazione del tempo, si associa una distribuzione continua sul tempo –$\psi_k \sim Beta$– a ogni topic.

2. Per ogni documento $d = 1, \ldots, D$:
   (a) Campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta_d \sim Dir_T(\alpha)$.
   (b) Per ogni parola w_{di}:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{di} \sim Mult(\theta_d)$.
      ii. Estraggo una parola dal topic corrispondente, $w_{di} \sim Mult(\phi_{z_{di}})$.
      iii. Estraggo un timestamp dalla distribuzione del tempo del topic corrispondente, $t_{di} \sim Beta(\psi_{z_{di}})$.

Il processo generativo (b) è preferibile per la generazione di un nuovo documento mai visto ed è

1. Per ogni topic $z = 1, \ldots, T$:
   (a) Campiono la distribuzione sulle V parole del topic da una distribuzione di Dirichlet simmetrica, $\phi_z \sim Dir_V(\beta)$.
   (b) Scelgo un insieme di parametri ψ_z per parametrizzare la distribuzione Beta.
2. Per ogni documento $d = 1, \ldots, D$:
   (a) Campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta_d \sim Dir_T(\alpha)$.
   (b) Estraggo un timestamp da una mistura delle distribuzioni Beta che ha come pesi i componenti delle proporzioni dei topic, $t_d \sim \sum_{z=1}^{T} \theta_{dz} Beta(\psi_z)$.
   (c) Per ogni parola w_{di}:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{di} \sim Mult(\theta_d)$.
      ii. Estraggo una parola dal topic corrispondente, $w_{di} \sim Mult(\phi_{z_{di}})$.

Questa formulazione permette prevedere il timestamp t^* –ovvero la collocazione nel tempo– di un documento a partire dalle parole presenti in esso: $t^* = \text{argmax}_t \prod_{i=1}^{N_d} \Pr(t|\psi_{z_i})$. Risulta inoltre interessante considerare la distribuzione dei topic condizionatamente ai timestamp in modo da poter osservare come si evolvono nel tempo le occorrenze dei topic; questa è ottenuta con la *regola di Bayes*: $\mathrm{E}[\theta_{z_i}|t] = \Pr(\theta_{z_i}|t) \propto \Pr(t|\theta_{z_i}) \Pr(z_i)$ dove $\Pr(z_i)$ può essere stimata dai dati oppure per semplicità assunta uniforme.

Dubey et al. 2013 propongono *nonparametric Topics over Time* (*npTOT*), un estensione del *TOT* che non fissa a priori il numero T di topic e permette di avere distribuzioni sul tempo con più picchi di probabilità[11] incorporando distribuzioni non parametriche su parole e timestamp. Il primo miglioramento si ottiene sostituendo la distribuzione di Dirichlet sui topic con un HDP (Teh et al. 2006) mentre il secondo sostituendo la distribuzione Beta

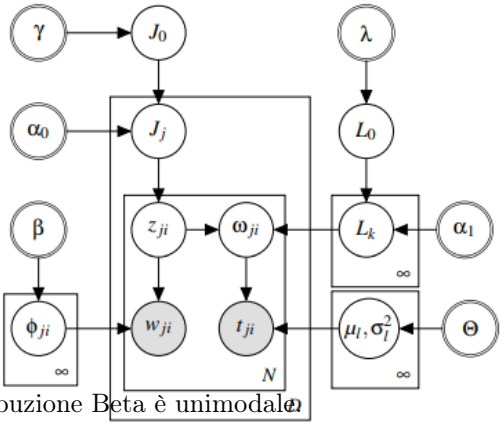


Figure 2: Plate diagram for nonparametric Topics over Time

[11] *TOT* permette di identificare un unico picco poiché la distribuzione Beta è unimodale

con una *Dirichlet process mixture of Gaussians*[12] in cui i pesi sono a loro volta estratti da un *HDP* in modo da avere i componenti della mistura condivisi tra i topic. Il processo generativo, rappresentato nell'immagine tratta da Dubey et al. 2013, è:

1. Campiono una distribuzione di base globale sulle proporzioni dei topic, $J_0|\gamma \sim \text{GEM}(\gamma)$.
2. Campiono una distribuzione di base globale sulle proporzioni delle componenti temporali, $L_0|\lambda \sim \text{GEM}(\lambda)$.
3. Per ogni topic $k = 1, 2, \ldots$:
   (a) Campiono la distribuzione sulle V parole del topic da una distribuzione di Dirichlet simmetrica, $\phi_k|\beta \sim Dir_V(\beta)$.
   (b) Campiono la distribuzione sulle componenti temporali del topic, $L_k|\alpha_1, L_0 \sim DP(\alpha_1, L_0)$.
4. Per ogni componente temporale $l = 1, 2, \ldots$:
   (a) Campiono una distribuzione sul tempo, $(\mu_t, \sigma_t^2)|\Theta \sim$ Normal-inverse Gamma(Θ).
5. Per ogni documento $j = 1, \ldots, D$:
   (a) Campiono le proporzioni dei topic, $J_j|\alpha_0, J_0 \sim DP(\alpha, J_0)$.
   (b) Per ogni parola $i = 1, \ldots, N_j$:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{ji}|J_j \sim J_j$.
      ii. Estraggo una parola dal topic corrispondente, $w_{ji}|\phi_{z_{ji}} \sim Mult(\phi_{z_{ji}})$.
      iii. Estraggo un timestamp, $t_{ji}|\mu_{w_{ji}}, \sigma^2_{w_{ji}} \sim N(\mu_{w_{ji}}, \sigma^2_{w_{ji}})$.

Riprendendo la metafora del *franchise di ristoranti cinesi* (*Chinese Restaurant Franchises*, *CRF*) esposta in Teh et al. 2006, il modello richiede due *CRF*: uno per il processo di Dirichlet gerarchico relativo alle parole, detto *word HDP*, e uno per quello relativo al tempo, detto *time HDP*. Nel primo ogni ristorante corrisponde a un documento e ogni piatto a un topic, nel secondo ogni ristorante a un topic e ogni piatto a una componente temporale, che a sua volta è associata a una distribuzione normale sul tempo. Infine, nella seguente immagine, sempre tratta da Dubey et al. 2013, si confrontano le distribuzioni sul tempo di due topic: si osserva che *npTOT* considera distribuzioni molto più flessibili che riescono a cogliere picchi di popolarità in istanti temporali differenti.

[12] Si veda: Daniel D. Walker, Kevin Seppi e Eric K. Ringger. «Topics over Nonparametric Time: A Supervised Topic Model Using Bayesian Nonparametric Density Estimation». In: *Proceedings of the Ninth UAI Conference on Bayesian Modeling Applications Workshop* - Volume 962. BMAW'12. Catalina Island, United States: CEUR-WS.org, 2012,pp. 74–83.

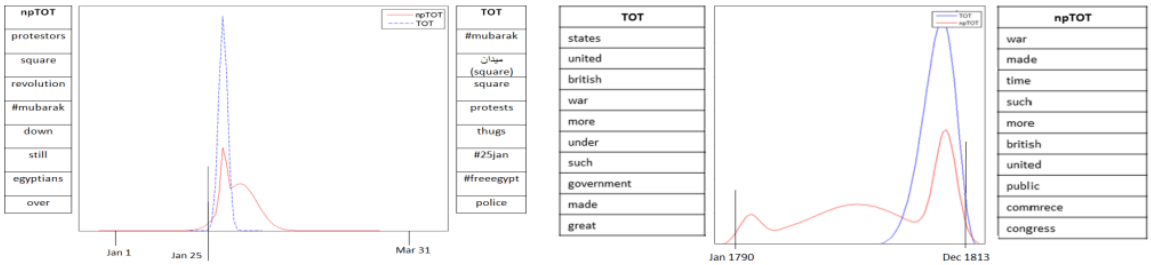
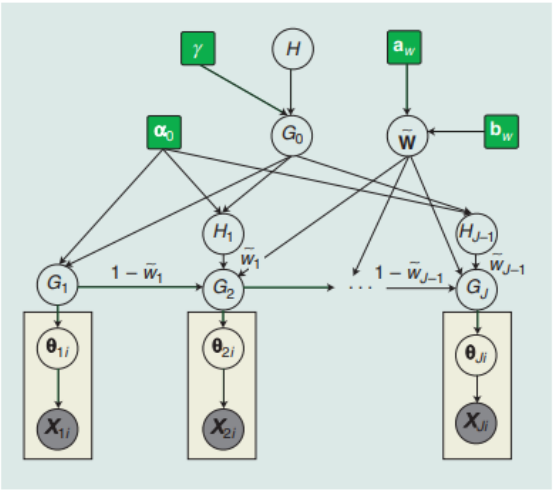


Figure 4: **Left:** Top eight most probable words in topics from TOT and npTOT corresponding to the Egyptian revolution (started on Jan 25) in Twitter dataset. **Right:** Top ten most probable words in topics from TOT and npTOT corresponding to conflicts involving the US and Britain in the State of the Union address dataset.

3.3.3 dHDP (2010)

D. Blei, Carin e Dunson 2010 propongono il *dynamic Hierarchical Dirichlet Process* (*dHDP*), un'estensione del *HDP* che modella l'evoluzione nel tempo delle proporzioni dei topic. Al posto di assegnare una distribuzione sui topic θ_d a ogni documento d, assegna una distribuzione sui topic G_t a tutti i documenti al tempo t; quindi G_t non indica più la struttura del singolo documento in termini di topic ma rappresenta la popolarità dei topic in un determinato periodo storico. Si assume che le G_t siano campionate come segue:



[FIG7] A graphical model of the dHDP model.

$$
\begin{aligned}
G_0 &\sim DP(\gamma, H) \\
\tilde{w}_t &\sim Beta(a_t, b_t) \\
H_t &\sim DP(\alpha_{0t}, G_0) \\
G_t &= (1 - \tilde{w}_t) G_{t-1} + \tilde{w}_t H_t
\end{aligned}
$$

con $\tilde{w}_1 = 1$, ovvero G_1 coincide con H_1. I parametri (a_t, b_t) sono tipicamente assunti indipendenti dal tempo per semplicità e selezionati in modo tale che $\tilde{w}_t$ abbia valori bassi con probabilità alta e meno frequentemente valori tendenti a 1, a cui corrisponde un'improvvisa variazione delle proporzioni dei topic al tempo t; in questo modo si ottiene un'evoluzione solitamente regolare che può avere raramente alcuni picchi. Nell'immagine tratta da D. Blei, Carin e Dunson 2010 (a_t, b_t) e α_{0t} sono assunti indipendenti dal tempo.

3.3.4 EvoHDP (2010)

Zhang et al. 2010 propongono il modello *Evolutionary Hierarchical Dirichlet Process* (*EvoHDP*) che estende *HDP* a uno scenario in cui si hanno più collezioni correlate varianti nel tempo (multiple correlated time-varying corpora). Queste vengono modellate con una serie di *HDP* con dipendenze temporali basate su un'assunzione Markoviana. Più nello specifico la dipendenza è espressa attraverso una mistura di due *DP*: il primo è quello dell'epoca precedente mentre il se-

condo, comune a tutte le collezioni ed epoche, è utilizzato per aggiornare il primo. Si considerino J collezioni su T epoche, siano

- J^t il numero di collezioni con documenti al tempo t;
- n_j^t il numero di documenti nella collezione j al tempo t;
- x_{ji}^t l'i-mo documento della collezione j al tempo t.

Si definiscono ora i *HDP*

- G è la misura di base condivisa dai *HDP* di tutte le epoche, detta *overall measure*;
- G_0^t è la misura globale all'epoca t, detta *snapshot global measure*;
- G_j^t è la misura locale per la j-ma collezione all'epoca t, detta *snapshot local measure*.

Il modello per generare x_{ji}^t è un modello mistura infinito

$$p_j^t(x|G_j^t) = \int G_j^t(\theta) f(x|\theta) d\theta$$

Per incorporare le dipendenze temporali tra epoche adiacenti vengono introdotti due tipi di dipendenze che modellano differenti comportamenti evolutivi:

- la differenza tra G_0^t e G_0^{t-1}, detta *global time dependency*, riflette l'evoluzione delle componenti globali in tutte le collezioni;
- la differenza tra G_j^t e G_j^{t-1}, detta *intra-corpus time dependency*, riflette l'evoluzione delle componenti all'interno della j-ma collezione.

Il processo generativo di *EvoHDP* è:

1. Campiono la *overall measure*, $G \sim DP(\xi, H)$.
2. Per ogni epoca $t = 1, \ldots, T$:
   a. Campiono la *snapshot global measure* all'epoca t da un DP in cui la misura di base è una mistura della *snapshot global measure* all'epoca $t-1$ e la *overall measure*, $G_0^t \sim DP(\gamma^t, w^t G_0^{t-1} + (1 - w^t)G)$.
   b. Per ogni collezione $j = 1, \ldots, J^t$:
      i. Campiono la *snapshot local measure* della collezione j all'epoca t da un DP in cui la misura di base è una mistura della *snapshot local measure* all'epoca $t-1$ e la *snapshot global measure* all'epoca t, $G_j^t \sim DP(\alpha^t, v_j^t G_j^{t-1} + (1 - v_j^t)G_0^t)$.
      ii. Per ogni documento $i = 1, \ldots, n_j^t$, estraggo il parametro θ_{ji}^t e l'osservazione x_{ji}^t attraverso un *hierarchical Dirichlet process mixture model*, $\theta_{ji}^t \sim G_j^t, \quad x_{ji}^t \sim F(x|\theta_{ji}^t)$.

La *stick-breaking construction*[13] di *EvoHDP* è

$$G = \sum_{k=1}^{\infty} \nu_k \delta_{\phi_k}, \qquad \boldsymbol{\nu} = (\nu_k)_{k=1}^{\infty} \sim \text{GEM}(\xi)$$

$$G_0^t = \sum_{k=1}^{\infty} \beta_k^t \delta_{\phi_k}, \qquad \boldsymbol{\beta}^t = (\beta_k^t)_{k=1}^{\infty} \sim DP(\gamma^t, w^t \boldsymbol{\beta}^{t-1} + (1 - w^t)\boldsymbol{\nu})$$

$$G_j^t = \sum_{k=1}^{\infty} \pi_{jk}^t \delta_{\phi_k}, \qquad \boldsymbol{\pi}_j^t = (\pi_{jk}^t)_{k=1}^{\infty} \sim DP(\alpha_0^t, v^t \boldsymbol{\pi}_j^{t-1} + (1 - v^t)\boldsymbol{\beta}^t)$$

per $j = 1, \ldots, J^t$, $t = 1, \ldots, T$. Si noti che gli *HDP* relativi a tutte le collezioni a tutte le epoche condividono lo stesso insieme infinito di componenti mistura, ovvero $\boldsymbol{\phi} = (\phi_k)_{k=1}^{\infty}$.

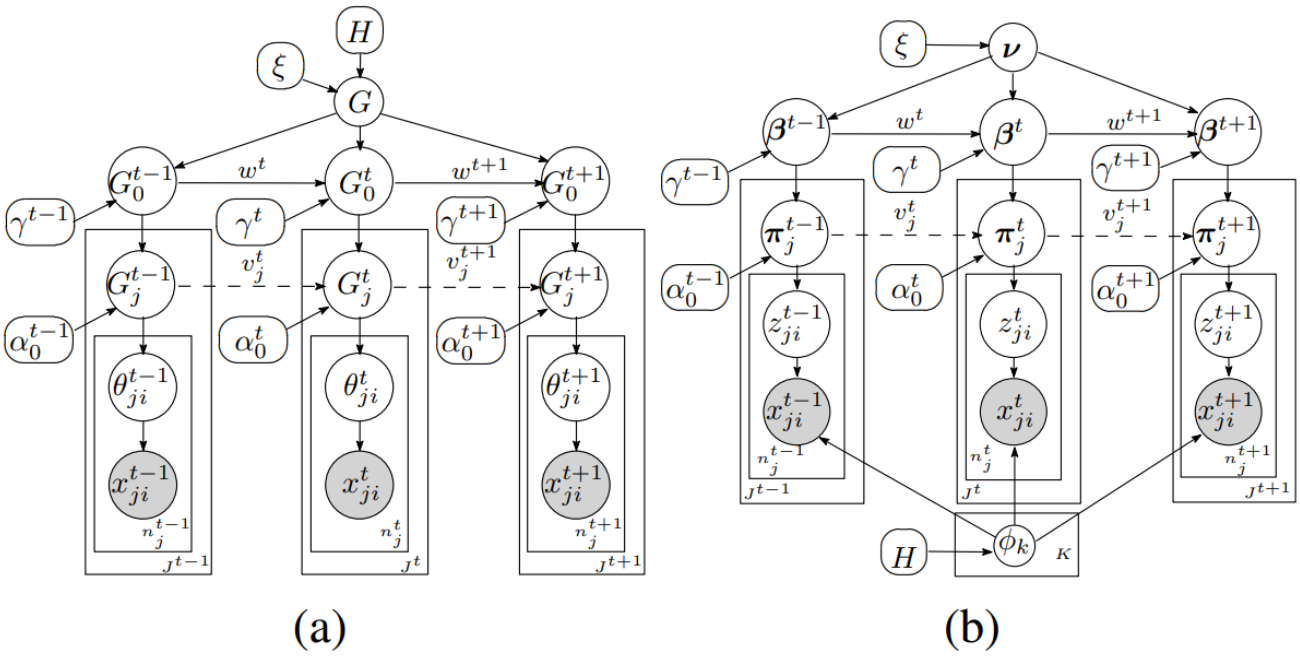


Figure 3: The graphical representation for the EvoHDP model. (a) The original representation. (b) The stick-breaking construction.

Si rimanda a Zhang et al. 2010 per la metafora del *chinese restaurant franchise*.

3.4 Utilizzo di metadati

In molte situazioni per ogni documento è possibile avere a disposizione delle informazioni aggiuntive come autore, titolo del documento, locazione geografica, link, hashtag, ... In generale, una qualsiasi informazione aggiuntiva assegnata ad un documento può essere considerata come un tag.

[13] Si veda l'Appendice A per un'introduzione alla rappresentazione di *DP* e *HDP* attraverso la *stick-breaking construction*.

3.4.1 Author-topic model (2004)

Rosen-Zvi et al. 2004 introducono l'*author-topic model* che estende la *LDA* includendo anche informazioni sull'autore. Assumendo che ogni documento sia stato scritto da uno o più autori, 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 di probabilità sulle parole propria di quel topic.

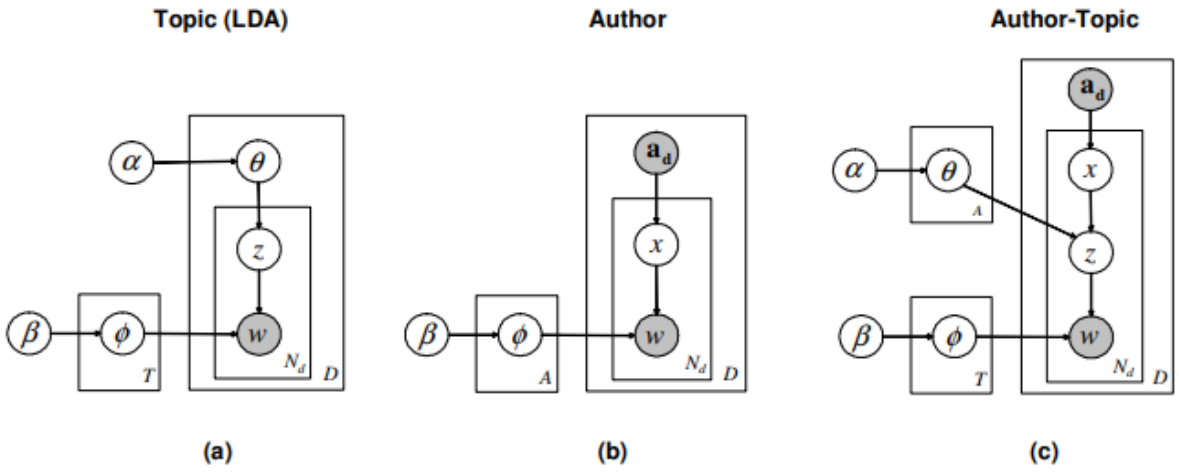


Figure 1: Generative models for documents. (a) Latent Dirichlet Allocation (LDA; Blei et al., 2003), a topic model. (b) An author model. (c) The author-topic model.

Nella figura tratta da Rosen-Zvi et al. 2004 si confrontano i processi generatori dell'*LDA*, l'*author model* –un modello che considera l'autore ma non topic latenti– e l'*author-topic model*. Il processo generativo dell'*author-topic model* è il seguente:

1. Campiono T distribuzioni sulle V parole, una per ogni topic, da una distribuzione di Dirichlet simmetrica, $\phi_t \sim Dir_V(\beta)$, $t = 1, \ldots, T$.
2. Campiono A distribuzioni sui T topic, una per ogni autore, da una distribuzione di Dirichlet simmetrica, $\theta_a \sim Dir_T(\alpha)$, $a = 1, \ldots, A$.
3. Per ogni parola n in ogni documento d:
   a. Estraggo un autore tra quelli associati al documento, $x_{d,n} \sim Mult\left(\left[\frac{1}{|\mathbf{a}_d|}, \ldots, \frac{1}{|\mathbf{a}_d|}\right]\right)$.
   b. Estraggo un topic dalle proporzioni dei topic dell'autore corrispondente, $z_{d,n}|x_{d,n}, \theta_{1:A} \sim Mult(\theta_{x_{d,n}})$.
   c. Estraggo una parola dal topic corrispondente, $w_{d,n}|z_{d,n}, \phi_{1:K} \sim Mult(\phi_{z_{d,n}})$.

Stimando i parametri θ_a e ϕ_t[14], si ottengono informazioni su quali topic l'autore a tende a scrivere e la rappresentazione del topic k in termini di parole. In pratica, al posto di avere la proporzione sui topic di ogni documento, si ha quella di ogni autore.

Si noti che un autore a assegnato a un documento d può essere visto come un tag che indica "l'autore a è associato al documento d": la stessa idea può essere usata quindi per qualsiasi tipo di dato generalizzando l'affermazione precedente in "la caratteristica a è associata al documento

[14]Si noti il cambio di notazione rispetto a D. M. Blei 2012: $\theta_a \sim Dir_T(\alpha)$ al posto di $\theta_d \sim Dir_K(\alpha)$ e $\phi_t \sim Dir_V(\beta)$ al posto di $\beta_k \sim Dir_V(\eta)$.

d". Sfruttando questa idea, Tsai 2011 propone il *tag-topic model*, un modello che coincide 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 documenti.

3.4.2 Labeled LDA (2009)

Ramage et al. 2009 introducono la *Labeled LDA*, un *topic model* in cui si impone una corrispondenza uno a uno tra topic latenti e le label assegnate ai documenti. Così facendo, si determinano le distribuzioni sulle parole di ogni topic come nella *LDA*, ma non è necessario attribuire un nome a ogni topic poiché il nome del topic è la label ad esso assegnata. Inoltre, le proporzioni dei topic indicano anche quanto le label sono influenti all'interno di ogni documento.

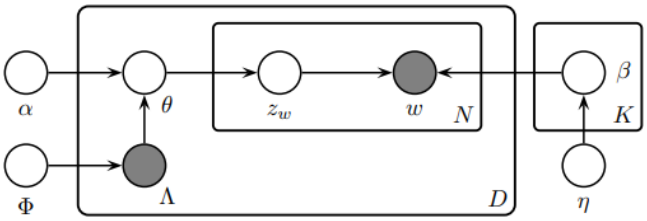


Figure 1: Graphical model of Labeled LDA: unlike standard LDA, both the label set $\mathbf{\Lambda}$ as well as the topic prior $\boldsymbol{\alpha}$ influence the topic mixture $\boldsymbol{\theta}$.

Il processo generativo della *Labeled LDA* è il seguente:

1. Campiono K distribuzioni sulle V parole, una per ogni topic/label, da una distribuzione di Dirichlet simmetrica, $\beta_k \sim Dir_V(\boldsymbol{\eta})$, $k = 1, \ldots, K$.
2. Per ogni documento $d = 1, \ldots, D$:
   a. Per ogni topic/label $k = 1, \ldots, K$:
      i. Genero l'indicatore della presenza del topic/label, $\Lambda_k^{(d)} \sim Bern(\mathbf{\Phi}_k)$.
   b. Genero la distribuzione a priori dei topic/label osservati, $\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha}$.
   c. Campiono le proporzioni dei topic/label osservati da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\theta}^{(d)} \sim Dir_{M_d}(\boldsymbol{\alpha}^{(d)})$.
   d. Per ogni parola $i = 1, \ldots, N_d$:
      i. Estraggo un topic/label dalle proporzione dei topic, $z_{d,i}|\boldsymbol{\theta}^{(d)} \sim Mult(\boldsymbol{\theta}^{(d)})$.
      ii. Estraggo una parola dal topic/label corrispondente, $w_{d,i}|z_{d,i}, \boldsymbol{\beta} \sim Mult(\boldsymbol{\beta}_{z_{d,i}})$.

Nel punto 2.a, la matrice $L^{(d)} \in \{0,1\}^{M_d \times K}$ seleziona le righe di $\boldsymbol{\alpha} \in (0,1)^{K \times V}$ corrispondenti ai topic/label osservati nel d-mo documento in modo da poter generare, nel punto 2.b, delle proporzioni non di tutti i topic/label (K) ma solo di quelli osservati (M_d). La matrice $L^{(d)}$ assume valori in $\{0,1\}$ e $L_{ij}^{(d)} = 1$ se l'i-mo topic/label del d-mo documento è il j-mo topic/label della collezione; quindi, ogni riga della matrice ha un solo valore non nullo e, in questo caso, se $L_{ij}^{(d)} = 1$, allora la i-ma riga di $\boldsymbol{\alpha}^{(d)}$ è la j-ma riga di $\boldsymbol{\alpha}$. Considero, ad esempio, un documento con $M_d = 2$ label in una collezione con $K = 4$ topic/label:

$$L^{(d)}\boldsymbol{\alpha} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} \alpha_2 \\ \alpha_3 \end{pmatrix}$$

Si noti che se le label assegnate ai documenti –ovvero $\Lambda_k^{(d)}$ $\forall d, k$– sono assunte note, la procedura di stima coincide essenzialmente con quella della *LDA*.

3.4.3 Twitter-LDA (2011)

W. X. Zhao et al. 2011 introducono *Twitter-LDA*, un modello creato ad-hoc per i *tweet* di *Twitter* in cui si assume che ogni utente abbia una sua sua proporzione dei topic, ogni *tweet* abbia un unico topic e ogni parola possa appartenere ad un topic o essere di sottofondo (*background words*): nel primo caso, prima si estrae un topic dalle proporzioni dei topic dell'autore e poi si estrae la parola dalla distribuzione sulle parole del topic; nel secondo caso, si estrae una parola dalla distribuzione sulle parole delle parole di sottofondo.

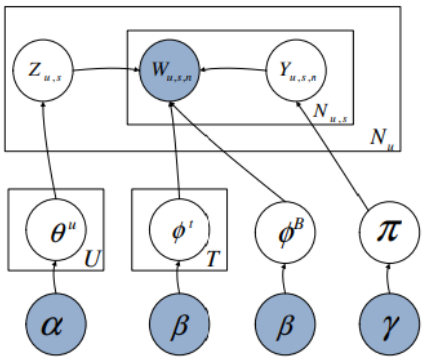


Fig. 2. Plate notation of our Twitter-LDA.

Il processo generativo del *Twitter-LDA* è il seguente:

1. Campiono la distribuzione sulle V parole delle parole di sottofondo, $\phi^{\mathcal{B}} \sim Dir_V(\beta)$. Campiono inoltre una *flag distribution*, $\pi \sim Dir_2(\gamma)$.
2. Campiono T distribuzioni sulle V parole, una per ogni topic, da una distribuzione di Dirichlet simmetrica, $\phi^t \sim Dir_V(\beta)$, $t = 1, \ldots, T$.
3. Per ogni utente $u = 1, \ldots, U$:
   a. Campione le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta^u \sim Dir_T(\alpha)$.
   b. Per ogni *tweet* $s = 1, \ldots, N_u$:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{u,s}|\theta^u \sim Mult(\theta^u)$.
      ii. Per ogni parola $n = 1, \ldots, N_{u,s}$:
         A. Seleziono l'origine della parola, $y_{u,s,n} \sim Bern(\pi)$.
         B. Se $y_{u,s,n} = 0$, estraggo una parola dalla distribuzione sulle V parole delle parole di sottofondo, $w_{u,s,n} \sim Mult(\phi^{\mathbf{B}})$; se $y_{u,s,n} = 1$, estraggo una parola dalla distribuzione sulle V parole del topic associato al documento, $w_{u,s,n} \sim Mult(\phi^{z_{u,s}})$.

3.4.4 Tag-Latent Dirichlet Allocation (TLDA, 2013)

Ma et al. 2013 introducono *Tag-Latent Dirichlet Allocation* (*TLDA*) che estende la *LDA* incorporando i tag osservati nel processo generativo. Il modello riprende la struttura dell'*author-topic model* (Rosen-Zvi et al. 2004), ma, al posto di selezionare casualmente un tag[15] tra quelli osservati, il modello assegna a ogni documento una proporzione dei tag e utilizza questa distribuzione per estrarre il tag di ogni parola.

Il processo generativo del *TLDA* è il seguente:

[15]Gli autori nell'*author-topic model* (Rosen-Zvi et al. 2004) possono essere considerati come tag assegnati a ogni documento.

1. Campiono L distribuzioni sui T topic, una per ogni tag, da una distribuzione di Dirichlet simmetrica, $\gamma_p \sim Dir_V(\rho)$, $p = 1, \ldots, L$.
2. Campiono T distribuzioni sulle V parole, una per ogni topic, da una distribuzione di Dirichlet simmetrica, $\beta_t \sim Dir_V(\phi)$, $t = 1, \ldots, T$.
3. Per ogni documento d:
   a. Campiono la distribuzione sui L_d tag osservati da una distribuzione di Dirichlet simmetrica, $\theta_d \sim Dir_{L_d}(\alpha)$.
   b. Per ogni parola $i = 1, \ldots, N_d$:
      i. Estraggo un tag tra quelli associati al documento, $e_{d,i} \sim Mult(\theta_d)$.
      ii. Estraggo un topic dalla distribuzione sui topic del tag corrispondente, $z_{d,i}|e_{d,i}, \gamma_{1:L} \sim Mult(\gamma_{e_{d,i}})$.
      iii. Estraggo una parola dal topic corrispondente, $w_{d,i}|z_{d,i}, \beta_{1:T} \sim Mult(\beta_{z_{d,i}})$.

Un tag può essere interpretato in due modi: visualizzando le parole più utilizzate dal topic più influente, se questo ha peso nettamente maggiore di tutti gli altri, oppure rappresentando graficamente la proporzione dei topic del tag. Inoltre, calcolando la *divergenza di Kullback–Leibler* o la *distanza di Hellinger* tra tutte le possibili coppie di tag, è possibile costruire una matrice $L \times L$ il cui elemento in posizione (i, j) indica la correlazione tra i tag i e j; anche in questo caso è possibile utilizzare rappresentazioni grafiche, ad esempio quelle utilizzate per le matrici di correlazione.

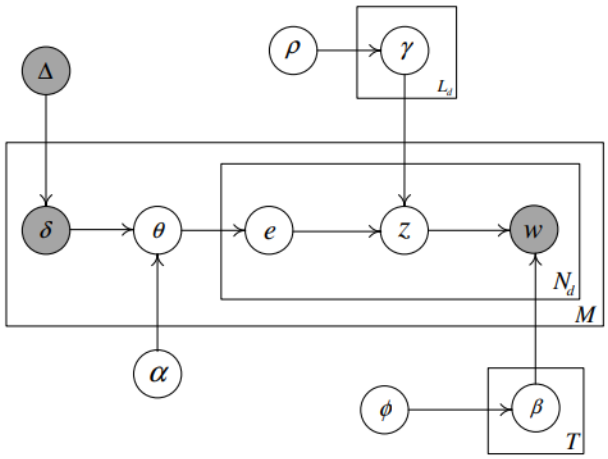


Figure 1. Graphical model of TLDA. Word w and tags δ are observed. Latent variables e and z are the tag and topic assignment to the word. Variables θ, γ, and β are latent variables. Tag set Δ is included so as to keep the completeness of the generative process.

3.4.5 Correlated Tag Learning (CTL, 2016)

Li et al. 2016 introducono *Correlated Tag Learning* (*CTL*) *model* che permette di cogliere la correlazioni tra tag attraverso un *processo di partecipazione normale logistico* (*logistic normal participation process*). Più nello specifico, il modello assume che in un documento i tag osservati influenzino la proporzione dei topic; l'influenza è modellata attraverso un vettore di partecipazione i cui componenti rappresentano l'importanza dei tag all'interno del documento. Il processo generativo del *CTL model* è il seguente:

1. Campiono K distribuzioni sulle V parole, una per ogni topic, da una distribuzione di Dirichlet simmetrica, $\psi_k \sim Dir_V(\pi)$, $k = 1, \ldots, K$.
2. Campiono L distribuzioni sui K topic, una per ogni tag, da una distribuzione di Dirichlet simmetrica, $\theta_t \sim Dir_V(\Lambda)$, $t = 1, \ldots, L$.

3. Per ogni documento $d = 1, \ldots, D$:
   a. Campiono η^d da una distribuzione normale multivariata, $\eta^d \sim N_L(\mu, \Sigma)$.
   b. Genero la matrice dei tag T^d a partire dalla lista dei tag osservati $\boldsymbol{t^d} = (t_1^d, \ldots, t_L^d)$, dove $T_{ij}^d \in \{0, 1\}$ è pari a 1 se l'i-mo tag del d-mo documento è il j-mo tag della collezione[16].
   c. Genero il vettore di partecipazione, $\varepsilon^d = \exp\{T^d \times (\eta^d)^\top\}$.
   d. Genero la proporzione sui topic come combinazione lineare delle distribuzioni dei tag, $\theta^d = \text{softmax}((\varepsilon^d)^\top \times T^d \times \boldsymbol{\theta}) = \frac{(\varepsilon^d)^\top \times T^d \times \boldsymbol{\theta}}{\sum_{i=1}^K [(\varepsilon^d)^\top \times T^d \times \boldsymbol{\theta}]_i}$.
   e. Per ogni parola $n = 1, \ldots, N$:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{dn}|\theta^d \sim Mult(\theta^d)$.
      ii. Estraggo una parola dal topic corrispondente, $w_{dn}|z_{dn}, \psi \sim Mult(\psi_{z_{dn}})$.

Nel punto 3.a, il vettore η^d tiene conto delle correlazioni tra i tag poiché μ rappresenta i peso medio dei vari tag all'interno di un documento e Σ la correlazione tra i tag. Nel punto 3.b, la matrice T^d seleziona le righe di η^d corrispondenti ai tag osservati nel documento d, mentre l'esponenziale assicura di avere solo valori positivi all'interno del vettore di partecipazione ε^d. Nel punto 3.c, come fatto per η^d, si selezionano le righe di $\boldsymbol{\theta}$ corrispondenti ai tag osservati e si combinano le distribuzioni sulle parole selezionate utilizzando i componenti del vettore di partecipazione come pesi; infine, la funzione *softmax* permette di ottenere un vettore i cui componenti sommano a 1. Sarebbe stato possibile generare i pesi da una distribuzione di Dirichlet, tuttavia in questo caso non si sarebbe tenuto conto della correlazione tra i tag. L'utilizzo di una distribuzione normale logistica rende più difficile l'inferenza sulla distribuzione a posteriori, ottenuta tramite *variational Expectation-Maximization algorith.* Si ottengono le stime di μ, Σ, ψ e Λ: a partire da Σ è possibile valutare le correlazioni tra i tag, mentre da ψ è possibile interpretare i topic identificati.

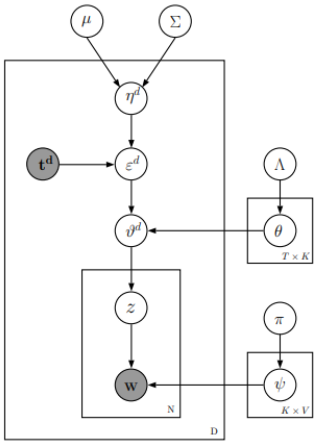


Figure 1: The graphical model of the CTL model, where each node denotes a random variable, a shaded node represents an observed variable, and edges indicate possible dependencies.

3.4.6 Hashtag-LDA (2016)

F. Zhao et al. 2016 introducono l'*Hashtag-LDA*, un modello basato sulla *LDA* che modella congiuntamente la relazione tra utenti, hashtag e parole nei microblog attraverso dei topic latenti. A ogni post è associato un unico topic estratto dalle proporzioni dei topic del suo autore, dalla distribuzione sulle parole del topic si generano separatamente le parole e gli hashtag: le prime sono generate come nella *LDA*, i secondi possono essere generati dalla proporzione sugli hashtag del topic associato al post oppure dalla proporzione degli hashtag globale, che è comune a tutti i

[16] Ogni riga T_i^d di T^d ha un solo valore non nullo: così facendo, ci si limita a selezionare e riordinare le righe di una matrice premoltiplicata per T^d; questa idea è tratta da Ramage et al. 2009.

topic e a tutti gli autori.
Il processo generativo dell'*Hashtag-LDA* è il seguente:

1. Campiono la distribuzione sui K hashtag globale da una distribuzione di Dirichlet simmetrica, $\omega \sim Dir_K(\lambda)$.
2. Campiono una *flag distribution*, $\omega \sim Dir_2(\rho)$.
3. Per ogni topic $t = 1, \ldots, T$:
   a. Campiono la distribuzione sui K hashtag da una distribuzione di Dirichlet simmetrica, $\pi_t \sim Dir_K(\gamma)$.
   b. Campiono la distribuzione sulle V parole da una distribuzione di Dirichlet simmetrica, $\phi_t \sim Dir_V(\beta)$.
4. Per ogni utente $u = 1, \ldots, U$:
   a. Campiono le proporzioni dei topic da una distribuzione di Dirichlet simmetrica, $\theta_u \sim Dir_T(\alpha)$.
   b. Per ogni microblog $n = 1, \ldots, N_u$:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{u,n}|\theta_u \sim Mult(\theta_u)$.
      ii. Per ogni parola $l = 1, \ldots, L_{u,n}$:
         - Estraggo una parola dal topic, $w_{u,n,l}|z_{u,n}, \phi_{1:T} \sim Mult(\phi_{z_{u,n}})$.
      iii. Per ogni hashtag $k = 1, \ldots, K_{u,n}$:
         A. Seleziono la distribuzione da cui estrarre l'hashtag, $g_{u,n,k}|\sigma \sim Bern(\sigma)$.
         B. Se $g_{u,n,k} = 0$, estraggo l'hashtag dalla distribuzione sui K hashtag globale, $h_{u,n,k}|\omega \sim Mult(\omega)$; se $g_{u,n,k} = 1$, estraggo l'hashtag dalla distribuzione sui K hashtag del topic, $h_{u,n,k}|z_{u,n}, \pi_{1:T} \sim Mult(\pi_{z_{u,n}})$.

Per ogni autore u, si ha una proporzione dei topic θ_u che indica su quali topic tende a scrivere. Per ogni topic t, si hanno due distribuzioni: una per le parole, ϕ_t, e una per gli hashtag, π_t, che indicano rispettivamente ciò che più utilizzato nei post associati a quel topic; solitamente le due distribuzioni sono molto simili ed entrambe interpretabili, tuttavia può capitare che per determinare il nome di un topic sia necessario guardare entrambe poiché una delle due distribuzioni non risulta abbastanza informativa.

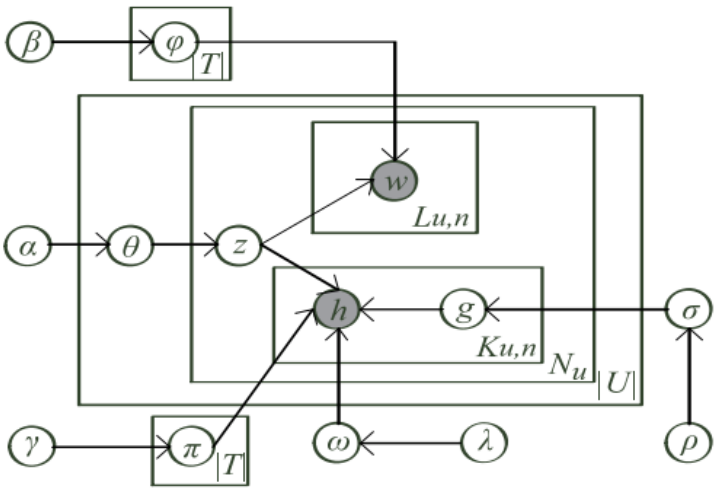


Fig. 2. Graphical representation of *Hashtag-LDA*.

3.5 Gestione della Sparsità

Si osserva 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, è 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 (Lin et al. 2014).

3.5.1 Dual-sparse Topic Model (DsparseTM, 2014)

Lin et al. 2014 introducono *Dual-sparse Topic Model* (*DsparseTM*), 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[17]; inoltre, introduce due coppie di distribuzioni a priori –*topic smoothing prior*, *weak topic smoothing prior* e *term smoothing prior*, *weak term smoothing prior*– per assicurare che le distribuzioni sui topic e sulle parole siano ben definite[18].

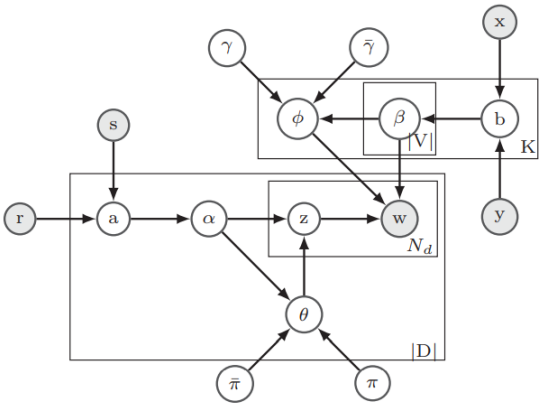


Figure 1: The graphical model of DsparseTM

Il processo generativo di *DsparseTM* è il seguente:

1. Per ogni topic $k = 1, \ldots, K$:
   a. Estraggo il parametro della Bernoulliana, $b_k \sim Beta(x, y)$.
   b. Per ogni termine $r = 1, \ldots, V$:
      i. Estraggo il *term selector*, $\beta_{kr} \sim Bern(b_k)$ con $\vec{\beta}_k = \{\beta_{kr}\}_{r=1}^{V}$.
      ii. Creo l'insieme dei *focused terms*, $B_k = \{r : \beta_{kr} = 1\}$.
      iii. Campiono la distribuzione sulle parole da una distribuzione di Dirichlet in cui si assegna una probabilità a priori maggiore ai *focused terms*, $\vec{\phi}_k \sim Dir_V(\gamma\ \vec{\beta}_k + \bar{\gamma}\ \vec{1})$.
2. Per ogni documento $d = 1, \ldots, D$:
   a. Estraggo il parametro della Bernoulliana, $a_d \sim Beta(s, t)$.
   b. Per ogni topic $k = 1, \ldots, K$:
      i. Estraggo il *topic selector*, $\alpha_{dk} \sim Bern(a_d)$ con $\vec{\alpha}_d = \{\alpha_{dk}\}_{k=1}^{K}$.
      ii. Creo l'insieme dei *focused topics*, $A_d = \{k : \alpha_{dk} = 1\}$.
   c. Campiono le proporzioni dei topic da una distribuzione di Dirichlet in cui si assegna una probabilità a priori maggiore ai *focused topics*, $\vec{\theta}_d \sim Dir_K(\pi\ \vec{\alpha}_d + \bar{\pi}\ \vec{1})$.
   d. Per ogni parola $i = 1, \ldots, N_d$:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{di} \sim Mult(\{\vec{\theta}_{dk} : k \in A_d\})$.
      ii. Estraggo una parola dal topic corrispondente, $w_{di} \sim Mult(\{\vec{\phi}_{z_{di}r} : r \in B_{z_{di}}\})$.

[17] Il modello è detto appunto *Dual-sparse* poiché gestisce la *Dual-Sparsity* nelle distribuzioni dei documenti e dei topic.

[18] Senza queste distribuzioni a priori, il *processo Spike and Slab* può generare distribuzioni in cui alcuni topic, o parole, non compaiono mai: oltre a complicate l'inferenza, ciò compromette la qualità dei topic.

Nel punto 1.b.iii, ogni componente del parametro della distribuzione di Dirichlet è pari a $\overline{\gamma}+\gamma$ se corrisponde a un *focused term*, vale $\overline{\gamma}$ altrimenti:

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

Si ha che la *weak smoothing prior* $\overline{\gamma}$ definisce la distribuzione a priori dei i termini non selezionati dal *term selector*; mentre, essendo $\overline{\gamma} \ll \gamma$, si ha che *smoothing prior* definisce distribuzione a priori dei termini selezionati, visto che $\overline{\gamma}$ risulta trascurabile nella somma $\overline{\gamma}+\gamma$. Inoltre, fissando $\overline{\gamma} \to 0$, si ha che i componenti di $\vec{\phi}_k$ corrispondenti ai valori non selezionati assumono valori talmente bassi –ma non nulli– da poter essere assunti pari a zero: è ragionevole quindi affermare che $\sum_{r \in B_k} \phi_{kr} = 1$. Riepilogando, nel punto 1.b.iii si campiona una distribuzione su tutte le V parole, ma con valori non trascurabili solo in corrispondenza dei *focused terms*: grazie a questa caratteristica del vettore, nel punto 2.d.ii è possibile considerare il vettore $|B_k| \times 1$ dei soli valori non trascurabili e comunque poter considerare il vettore come un vettore di probabilità.

Lo stesso ragionamento vale per i *focused topic* ai punti 2.c e 2.d.i: 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 per introdurre sparsità nel modello.

3.6 Correlazione

3.6.1 Correlated Topic Model (CTM, 2007)

D. M. Blei e Lafferty 2007 propongono il *Correlated Topic Model* (CTM), una variazione della *LDA* che modella direttamente la correlazione tra topic, in cui la distribuzione di Dirichlet per generare le proporzioni dei topic di ogni documento è sostituita da una distribuzione normale logistica. La distribuzione di Dirichlet genera vettori con componenti quasi del tutto indipendenti[19]: al contrario, la distribuzione normale logistica permette generare vettori la cui correlazione tra i componenti è determinata dalla matrice di varianza e covarianza.
Il processo generativo del CTM è il seguente:

1. Campiono K distribuzioni sulle V parole, una per ogni topic, da una distribuzione di Dirichlet simmetrica, $\beta_k \sim Dir_V(\eta)$, $k = 1, \ldots, K$.
2. Per ogni documento $d = 1, \ldots, D$:
   a. Campiono le proporzioni dei topic da una distribuzione normale logistica, $\eta_d \sim N(\mu, \Sigma)$, $\theta_d = \text{softmax}(\eta_d)$.
   b. Per ogni parola $n = 1, \ldots, N$:
      i. Estraggo un topic dalle proporzioni dei topic, $z_{d,n}|\theta_d \sim Mult(\theta_d)$.
      ii. Estraggo una parola dal topic corrispondente, $w_{d,n}|z_{d,n}, \beta_{1:K} \sim Mult(\beta_{z_{d,n}})$.

[19] I componenti di un vettore generato da una distribuzione di Dirichlet hanno una leggera correlazione negativa poiché devono sommare a 1.

Per ogni documento d, prima si genera un vettore da una distribuzione normale la cui media μ rappresenta il peso medio dei vari topic e la matrice di varianza e covarianza Σ la correlazione tra i topic, poi si utilizza la funzione *softmax* per mappare il vettore sul simplesso dei topic, ovvero renderlo un vettore di probabilità.

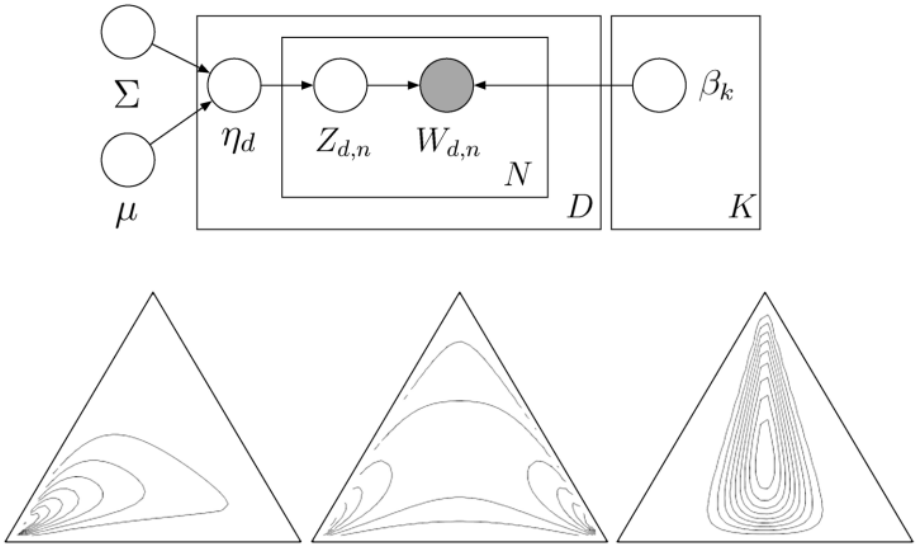


FIG. 1. Top: *Probabilistic graphical model representation of the correlated topic model. The logistic normal distribution, used to model the latent topic proportions of a document, can represent correlations between topics that are impossible to capture using a Dirichlet.* Bottom: *Example densities of the logistic normal on the 2-simplex. From left: diagonal covariance and nonzero-mean, negative correlation between topics 1 and 2, positive correlation between topics 1 and 2.*

4 Altri modelli

4.1 Multiscale Topic Tomography Model (MTTM, 2007)

Nallapati et al. 2007 propongono il *Multiscale Topic Tomography Model* (*MTTM*), un nuovo approccio che si basa su processi di Poisson non-omogenei e *multi-scale Haar wavelet analysis*; i primi sono utilizzati come processo generativo di un collezione suddivisa in epoche mentre il secondo permette di legare tra loro i parametri di processi relativi a diverse epoche. Assumendo che la collezione sia ordinata e divisa in 2^S chunk, detti *epoche*, ognuno di numerosità M; assumendo inoltre che a ogni epoca t sia associato un parametro $\boldsymbol{\mu}_t \in \mathbb{R}^{V \times K}$ il cui componente μ_{tkw} indica il conteggio atteso della parola w dal topic k durante l'epoca t mentre la k-ma colonna $\boldsymbol{\mu}_{tk}$ è la distribuzione delle parole dal topic k durante l'epoca t. Il processo generativo è:

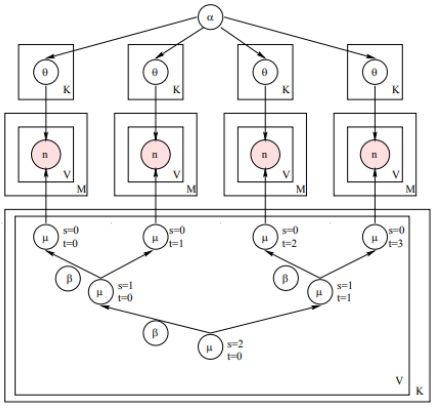


Figure 4: Graphical representation of MTTM for S=2: we purposely omitted the hyper-parameters in the figure for clarity.

Per ogni epoca $t = 0, \dots, 2^S - 1$:

1. Per ogni topic $k = 1, \dots, K$:
   a. Genero i pesi del k-mo topic nei documenti da una distribuzione di Dirichlet simmetrica, $\boldsymbol{\theta}_{tk} = \{\theta_{tk1}, \dots, \theta_{tkM}\} \sim Dir_M(\alpha)$.
2. Per ogni documento $d = 1, \dots, M$:
   a. Per ogni parola $w = 1, \dots, V$:
      i. Genero il conteggio, $n_{tdw} \sim Poi(\sum_k \theta_{tkd}\mu_{tkw})$.

Si noti che i pesi dei K topic non sommano più a 1 all'interno di ogni documento come nella *LDA* ma, fissato un topic k, sommano a 1 all'interno dei documenti di un'epoca[20]; l'idea di base è che a ogni epoca si ha una emissione di un topic e questa viene suddivisa tra i documenti di quell'epoca, infatti θ_{tkd} indica quanto il d-mo documento dell'epoca t cattura il topic k e sono tali che $\sum_{d=1}^{M} \theta_{tkd} = 1$ per t e k fissati. Per legare i parametri $\boldsymbol{\mu}_t$ di diverse epoche, si definiscono dei *multiscale wavelet parameters* dati dal seguente albero binario:

$$\begin{aligned} \boldsymbol{\mu}_t^{(S)} &= \boldsymbol{\mu}_t && \text{per } t = 0, \dots, 2^S - 1 \\ \boldsymbol{\mu}_t^{(s)} &= \boldsymbol{\mu}_{(2t)}^{s+1} + \boldsymbol{\mu}_{(2t+1)}^{s+1} && \text{per } s = 0, \dots, S-1 \text{ e } t = 0, \dots, 2^s - 1 \end{aligned}$$

dove s, detto *scale*, è la profondità dell'albero. Si ha quindi che ogni nodo foglia ($s = S$) corrisponde ad un'epoca e i nodi non foglia ($0 \leq s \leq S-1$) corrispondono a un intervallo temporale più ampio che include tutte le epoche legate ai nodi figli. Si definisce inoltre il *canonical multiscale parameter* $\beta_t^{(s)}$ che indica come il parametro $\boldsymbol{\mu}_t^{(s)}$ viene suddiviso tra i nodi figli ed è definito come il rapporto tra il nodo figlio sinistro e il nodo padre:

$$\beta_t^{(s)} = \frac{\mu_{(2t)}^{(s+1)}}{\mu_t^{(s)}} \quad \text{per } s = 0, \dots, S-1 \text{ e } t = 0, \dots, 2^S - 1$$

Il processo generativo per i parametri di Poisson è:

Per ogni topic $k = 1, \dots, K$:

1. Per ogni parola $w = 1, \dots, V$:
   a. Genero $\mu_{0kw}^{(0)} \sim Gamma(\lambda_\mu, \delta_\mu)$.
   b. Per ogni *scale* $s = 0, \dots, S-1$:
      i. Per ogni epoca $t = 0, \dots, 2^S - 1$:
         - Genero $\beta_{tkw}^{(s)} \sim Beta(\delta_\beta, \delta_\beta)$

Per stimare i parametri a posteriori si utilizza inferenza variazionale e si introducono due parametri variazionali: una quantità proporzionale alla probabilità a posteriori che il d-mo documento dell'epoca t catturi il k-mo topic γ_{tkd} e la probabilità a posteriori che la parola w nel d-mo documento dell'epoca t provenga dal k-mo topic ϕ_{tdwk}. Considerando le immagini di Nallapati et al. 2007, *MTTM* permette di:

[20] Questo è un aspetto negativo del modello poiché non è più possibile fare inferenza su un singolo documento, bisogna sempre considerare tutti i documenti di un'epoca.

- valutare l'evoluzione dei contenuti di un topic al variare delle epoche grazie alla rappresentazione ad albero binario dei parametri; nell'articolo è detta *zoom feature* e consiste nell'identificare le parole con valori più alti in $\boldsymbol{\mu}_{tk}^{(s)}$ al variare di s. Figure 5.

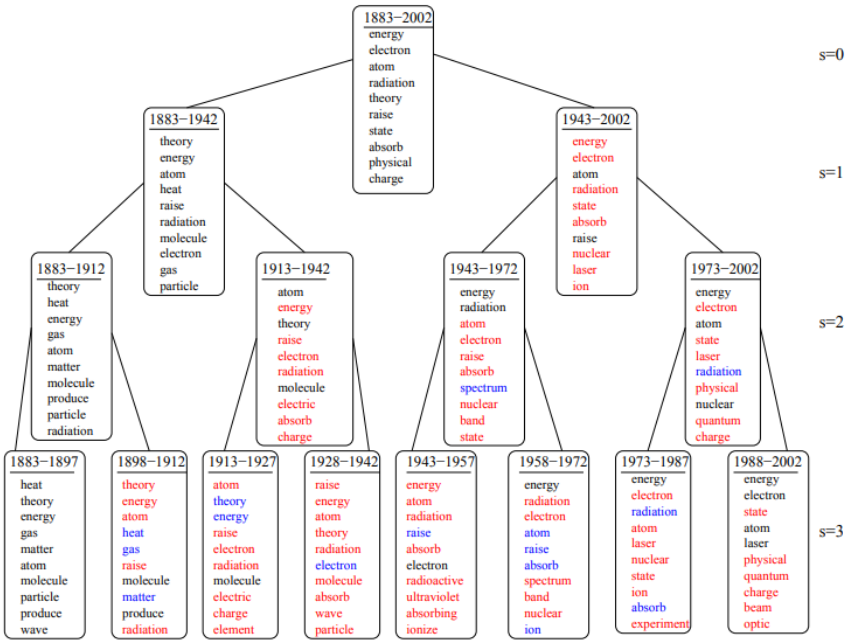


Figure 5: **A 4-scale representation of a topic which we manually labeled "Particle physics": best seen in color. Words colored red are those whose relative importance in the topic has gone up compared to previous epoch at the same scale. Words colored blue are those whose relative importance has gone down. Words not colored have retained their position compared to previous epoch.**

- valutare l'importanza di una parola in un determinato topic al variare delle epoche; l'importanza è data dal conteggio atteso della parola w dal topic k durante l'epoca t, μ_{tkw}, quindi si traccia il grafico di $\mu_{1kw}, \ldots, \mu_{2^S kw}$ per k e w fissati. Figure 7.
- valutare la frequenza di un topic al variare delle epoche in modo da poter identificare i periodi di maggior e minor popolarità di ogni topic; la frequenza del k-mo topic nell'epoca t è data da $\gamma_{tk} = \sum_{d=1}^{M} \gamma_{tkd}$. Figure 9.

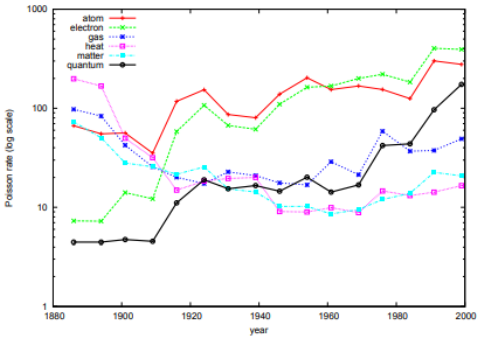


Figure 7: **Evolution of content bearing words in the topic manually labeled as "Particle physics": the words "atom", "electron" and "quantum" gain prominence with time, while words such as "heat" and "gas" lose ground, indicating a paradigm shift in the field from macro-matter to micro-matter.**

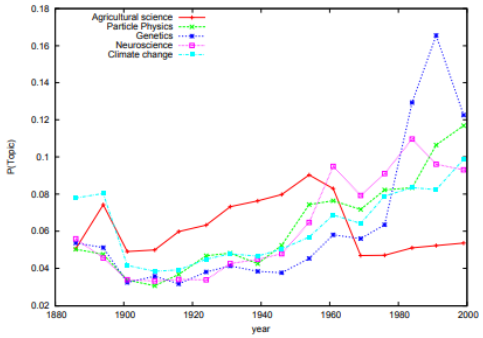


Figure 9: **Occurrence probability of topics with time: we plotted the statistic γ_{tk} for the two topics we analyzed earlier, namely "particle physics" and "genetics" and for three other topics, which we identified as "agricultural science", "neuroscience" and "climate change". The plot reveals some interesting patterns. For example, while *agricultural science* remains more or less stable with time, we see an explosion of *genetics* in the 1990's. The topics of *climate change*, *atomic physics* and *neuroscience* also exhibit an increasing prominence in the late 20th century, consistent with the trends in the real-world.**

4.2 Embedded Topic Model (ETM, 2019)

Dieng, F. J. R. Ruiz e D. M. Blei 2020 introducono un modello generativo per documenti che combina i *topic model* e i *word embeddings*, detto *Embedded Topic Model* (*ETM*). Il modello risolve una limitazione della *LDA*: non si riesce ad ottenere un buon modello –interpretabile e con una buona capacità predittiva– quando il vocabolario è troppo grande. *ETM* permette di considerare sia le parole rare sie le *stop words*. Inoltre, mantiene gli aspetti interessanti di entrambi i metodi: fornisce una struttura semantica latente interpretabile dei documenti (*LDA*) e fornisce una rappresentazione a bassa dimensionalità delle parole e dei topic (*word embeddings*).

Come nella *LDA*, ogni documento è una mistura finita di topic e ogni parola osservata è assegnata ad un topic, tuttavia la probabilità condizionata di ogni topic ha una forma log-lineare che involve una rappresentazione a bassa dimensionalità del vocabolario. In particolare, la distribuzione sulle parole di un topic è proporzionale all'esponenziale del prodotto interno dell'embedding del topic $\alpha_z \in \mathbb{R}^L$ e gli embedding delle parole $\rho_{1:V}$.

Si consideri una collezione di D documenti e un vocabolario di V parole, sia $\rho_{1:V} = (\rho_1 \dots \rho_V)^\top \in \mathbb{R}^{L \times V}$ l'embedding matrix, con ρ_v *word embedding* della v-ma parola, e sia $\alpha_{1:K} = (\alpha_1 \dots \alpha_K)^\top \in \mathbb{R}^{L \times K}$ la topic embedding matrix, con α_k *word embedding* del k-mo topic. La dimensione L degli *embedding* è fissata a priori; rappresentare sia le parole sia i topic come punti L-dimensionali di un stesso *embedded space* permette al modello di essere robusto alla presenza di *stop words* e parole rare.
Il processo generativo dell'*ETM* è il seguente:

1. Campiono D proporzioni dei topic, una per ogni documento, da una distribuzione normale logistica di media nulla e a componenti indipendenti con varianza unitaria, $\theta_d \sim LN(0, I_K)$.
2. Per ogni parola n in ogni documento d:
   a. Estraggo un topic dalla proporzione dei topic, $z_{dn}|\theta_d \sim Mult(\theta_d)$.
   b. Estraggo una parola dal topic corrispondente, $w_{dn} \sim Mult(\text{softmax}(\rho_{1:V}^\top \alpha_{z_{dn}}))$.

Si osservano due differenze rispetto alla *LDA*: nel punto 1, si usa la distribuzione logistico-normale per motivi d'inferenza; un vettore è generato come $\delta_d \sim N_K(0, I_k)$, $\theta_d = \text{softmax}(\delta_d)$; nel punto 2.b, si riprende l'idea del *Continuous Bag of Words* (*CBOW*) *model*, ma, al posto di considerare le parole vicine, si considera il contesto della parola, cioè il topic assegnato ad essa. La distribuzione sulle V parole del k-mo topic è data da $\beta_k = \text{softmax}(\rho_{1:V}^\top \alpha_k) \in [0, 1]^V$.

Per effettuare l'inferenza è necessario ricorrere all'*inferenza variazionale*; in particolare, è possibile considerare due casi: nel primo, detto *ETM*, si stimano gli *embedding* sia delle parole sia dei topic; nel secondo, detto *ETM-PWE*, si stimano solo gli *embedding* dei topic. Più nello specifico, nel secondo caso gli *embedding* delle parole sono assunti noti –sono già stati calcolati con altri metodi– e quindi i topic sono determinati da un *embedding* space già fissato.

5 Word Embeddings

5.1 Exponential Family EMBedding (EF-EMB)

Gli *Exponential Family EMBedding* (*EF-EMB* o *EFE*), proposti in Rudolph, F. J. R. Ruiz et al. 2016, sono una classe di metodi che estendono l'idea dei *word embeddings* a anche altri tipi di dati ad alta dimensionalità. Si hanno tre ingredienti:

1. una *context function* che definisce il contesto c_i di ogni osservazione $x_i \in \mathbb{R}^D$; c_i è un insieme di indici diversi da i mentre $\mathbf{x}_{c_i}$ è una matrice formata dai vettori delle osservazioni appartenenti al contesto.
2. una *conditional exponential family* tramite cui il modello *EF-EMB* modella ogni osservazione x_i condizionatamente al suo contesto $\mathbf{x}_{c_i}$:

   $$x_i|\mathbf{x}_{c_i} \sim \text{ExpFam}(\eta_i \mathbf{x}_{c_i}, t(x_i))$$

   dove $t(x_i)$ la statistica sufficiente e $\mathbf{x}_{c_i}$ è un parametro naturale, funzione a sua volta di due parametri vettoriali latenti:

   - l'*embedding vector* $\rho[i] \in \mathbb{R}^{K\times D}$ che aiuta a governare la distribuzione dell'i-ma osservazione x_i;
   - il *context vector* $\alpha[i] \in \mathbb{R}^{K\times D}$ che aiuta a governare la distribuzione delle osservazioni in cui x_i fa parte del *contesto*.

   Nell'articolo si considera il caso particolare dei *linear embedding* in cui $\eta_i \mathbf{x}_{c_i}$ è una funzione di una combinazione lineare dei vettori latenti:

   $$\eta_i \mathbf{x}_{c_i} = f_i\left(\rho[i]^\top \sum_{j\in c_i} \alpha[j] x_j\right)$$

   L'idea è quindi che la probabilità di vedere una particolare osservazione x_i dipende da due vettori latenti, che forniscono rispettivamente informazioni riguardo l'osservazione stessa e quelle vicine.
3. una *embedding structure* che determina come i vettori latenti sono condivisi tra le osservazioni; attraverso questa condivisione tra osservazioni differenti è possibile ottenere gli *embedding*. In ambito testuale i parametri sono condivisi da tutte le osservazioni, $\rho[i] = \rho$ e $\alpha[i] = \alpha$ $\forall i$, ovvero i vettori latenti non dipendono dalla posizione della parola all'interno del testo.

Per stimare i vettori latenti $\boldsymbol{\rho} = (\rho[1], \ldots, \rho[I])^\top$ e $\boldsymbol{\alpha} = (\alpha[1], \ldots, \alpha[I])^\top$ si minimizza la seguente funzione obiettivo rispetto a $\boldsymbol{\rho}$ e $\boldsymbol{\alpha}$:

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = \sum_{i=1}^{I} (\eta_i^\top t(x_i) - a(\eta_i)) + \log p(\boldsymbol{\rho}) + \log p(\boldsymbol{\alpha})$$

5.2 Dynamic Bernoulli EMBeddings (D-EMB)

Rudolph e D. Blei 2018 propongono i *Dynamic Bernoulli EMBeddings* (*D-EMB*), un'estensione dei *EF-EMB*[21], e li utilizzano per analizzare collezioni di documenti testuali scritti in intervalli

[21] In questo caso, il contesto sono le parole vicine a quella d'interesse e la famiglia esponenziale è una Bernoulliana.

temporali differenti; in particolare, dividono i documenti in diversi intervalli temporali e calcolano i *word embedding* per ogni intervallo differenti in modo da poter visualizzare l'evoluzione del significato delle parole. L'evoluzione è modellata attraverso un *Random Walk Gaussiano* in modo da non avere il problema dell'allineamento[22].

Un corpus è considerato come una sequenza di N termini, $(x_1, \dots, x_N)$, da un vocabolario di V parole, ogni termine $x_i \in \{0,1\}^V$ è un vettore indicatore con un valore non nullo al v-mo componente se l'i-mo termine del corpus è la v-ma parola del vocabolario. La distribuzione condizionata di $x_{iv} \in \{0,1\}$ è

$$x_{iv}|\mathbf{x}_{c_i} \sim Bern(p_{iv})$$

dove $\mathbf{x}_{c_i}$ sono le parole vicine a x_i, ovvero il contesto, e la probabilità p_{iv} è definita come

$$p_{iv} = \frac{e^{\eta_{iv}}}{1 + e^{\eta_{iv}}} \qquad \text{con} \qquad \eta_{iv} = \rho_v^{(t_i)\top} \left(\sum_{j \in c_i} \sum_{v'} \alpha_{v'} x_{jv'} \right)$$

dove $\alpha_v \in \mathbb{R}^K$ non dipende dalla posizione i della parola all'interno dell'intero corpus mentre $\rho_v^{t_i} \in \mathbb{R}^K$ non dipende dalla posizione i sono all'interno dello stesso intervallo temporale. Inoltre, si usa un *RW Gaussiano* come distribuzione a priori dei vettori latenti:

$$\begin{aligned} \alpha_v, \rho_v^{(0)} &\sim N(0, \lambda_0^{-1} I_{2K}) \\ \rho_v^{(t)} &\sim N(\rho_v^{(t-1)}, \lambda^{-1} I_K) \end{aligned}$$

Si hanno quindi $\alpha_1, \dots, \alpha_V$ condivisi da tutte le posizioni i del corpus e da tutti gli intervalli temporali, mentre $\rho_1^{t_i}, \dots, \rho_V^{t_i}$ sono condivisi solo dalle posizioni all'interno dell'intervallo temporale t_i. Assumendo di avere T intervalli temporali, si avranno V *context vector* e VT *embedding vector*, quindi per ogni parola si avrà un unico *context vector* e T *embedding vector*, uno per ogni intervallo temporale: a partire da quest'ultimi è possibile valutare come si evolve il significato di una parola v sia osservando direttamente come varia l'*embedding vector* $\rho_v^{t_i}$ al variare di t_i sia osservando come variano le 10 parole più simili nei vari intervalli temporali.

5.3 Structured Exponential Family Embeddings (S-EFE)

Rudolph, F. Ruiz et al. 2017 propongono i *Structured Exponential Family Embeddings (S-EFE)*, un'estensione dei *EF-EMB* per dati raggruppati in cui per ogni oggetto si vuole stimare un diverso *embedding* per ogni gruppo. L'articolo si concentra sulle parole di un corpus e quindi ha come obiettivo determinare come il significato di una stessa parola varia all'interno di diversi gruppi di documenti. Come in *D-EMB* proposto da Rudolph e D. Blei 2018, il nuovo modello è costruito imponendo una particolare *sharing structure*: i *context vector* sono condivisi tra tutte le posizioni i del corpus come nel modello originale e tra tutti i gruppi, $\alpha_v[i] = \alpha_v \ \forall i$, mentre gli *embedding vector* sono condivisi solo dalle posizioni all'interno dello stesso gruppo, $\rho_v[i] = \rho_v^{(s_i)}$ $\forall i$. Assumendo di avere S gruppi, si avranno V *context vector* e VS *embedding vector*, quindi per ogni parola si avrà un unico *context vector* e S *embedding vector*, uno per ogni gruppo.

[22]Si ha il problema dell'allineamento quando si stima un modello differente per ogni intervallo temporale e poi si cerca di confrontare i vettori ottenuti in modelli diversi.

Per ottenere buoni *embedding vector* e per gestire un eventuale carenza di parole in alcuni gruppi, vengono proposti due metodi per legare tra loro i vari ρ_v^s. In entrambi i casi l'idea è che esiste una insieme di *global embedding* $\rho_1^{(0)}, \dots, \rho_V^{(0)}$ da cui è possibile generare quelli propri di ogni gruppo $\rho_1^{(s)}, \dots, \rho_V^{(s)}$ $\forall s$:

- *Hierarchical embedding structure*
  Si impone una struttura gerarchica in cui si assume $\rho_v^{(s)} \sim N(\rho_v^{(0)}, \sigma_\rho^2 I)$, σ^2 iperparametro fissato. In questo caso non si riduce il numero di parametri su cui fare inferenza ma si legano tra loro attraverso una distribuzione a priori comune.
- *Amortization*
  Si assume che i *per-group embedding* siano il risultato di una funzione deterministica dei *global embedding*, $\rho_v^{(s)} = f_s(\rho_v^{(0)})$; in particolare, ogni gruppo ha una diversa funzione $f_s(\cdot)$ che viene parametrizzata attraverso una *rete neurale*. Siano $\phi^{(s)} \in \mathbb{R}^P$ i parametri della *rete neurale* stimata per il gruppo s, allora $\rho_v^{(s)}$ si ottiene come $\rho_v^{(s)} = f_s(\rho_v^{(0)}) = f_s(\rho_v^{(0)}, \phi^{(s)})$. Questo metodo permette di passare da $KL(S+1)$ parametri a $2KL + SP$, dove K è la dimensione degli *embedding vector* e L il numero di oggetti; essendo solitamente $L \gg P$, si una significativa riduzione dei parametri.

A Dirichlet Processess & Hierarchical Dirichlet Processess

Si consideri il problema in cui le osservazioni sono organizzate in gruppi; assumo inoltre che ci sia scambiabilità sia tra i gruppi sia entro i gruppi, ovvero

- le osservazioni $x_{j1}, x_{j2}, \ldots$ sono scambiabili all'interno del gruppo j, ovvero per j fissato;
- i gruppi $\mathbf{x}_1 = (x_{11}, x_{12}, \ldots), \mathbf{x}_2 = (x_{21}, x_{22}, \ldots), \ldots$ sono scambiabili.

Si consideri il seguente modello probabilistico:

$$\theta_{ji}|G_j \sim G_j, \qquad x_{ji}|\theta_{ji} \sim F(\theta_{ji}), \qquad \forall j, i$$

dove il parametro θ_{ji}, detto *fattore*, specifica il componente mistura associato all'osservazione x_{ji}, $F(\theta_{ji})$ è la distribuzione a priori per i fattori $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots)$ associati al gruppo j; si assume che i fattori sia condizionatamente indipendenti data la distribuzione G_j.

A.1 Dirichlet Process

Un *processo di Dirichlet* $DP(\alpha_0, G_0)$ è definito come la distribuzione di una misura di probabilità G sullo spazio misurabile $(\Theta, \mathcal{B})$ tale che, per qualsiasi partizione finita $(A_1, \ldots, A_r)$ di Θ, il vettore di probabilità $(G(A_1), \ldots, G(A_r))$ è distribuito come una distribuzione di Dirichlet finito-dimensionale di parametri $(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r))$, ovvero

$$(G(A_1), \ldots, G(A_r)) \sim Dir_r(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r))$$

A.1.1 The Stick-Breaking Construction

Misure estratte da una DP sono discrete con probabilità 1; questa proprietà può essere dimostrata con la *stick-breaking construction* che si basa su sequenze i.i.d. di variabili casuali $(\pi'_k)_{k=1}^{\infty}$ e $(\phi_k)_{k=1}^{\infty}$:

$$\pi'_k|\alpha_0, G_0 \sim Beta(1, \alpha_0)$$
$$\phi_k|\alpha_0, G_0 \sim G_0$$

È stato dimostrato[23] che G definita nel seguente modo è una $DP(\alpha_0, G_0)$:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

dove δ_ϕ è una misura di probabilità concentrata in ϕ –ovvero $\Pr(\delta_\phi = \phi) = 1$, 0 altrimenti– e $\pi_k = \pi'_k \prod_{l=1}^{k-1}(1 - \pi'_l)$ è tale che $\sum_{k=1}^{\infty} \pi_k = 1$ con probabilità 1; $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ può essere quindi interpretata come una misura di probabilità su interi positivi e si scrive $\boldsymbol{\pi} \sim \text{GEM}(\alpha_0)$.

[23] La dimostrazione è riportata in: Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

A.1.2 The Chinese Restaurant Process

Una seconda prospettiva del DP è fornita dal *Pólya urn scheme* che mostra che estrazioni da una DP sono discrete ed esibiscono una proprietà di clustering. Sia $\theta_1, \theta_2, \ldots \overset{iid}{\sim} G$ sequenza di variabili casuali condizionatamente indipendenti dato G –e quindi scambiabili–, la distribuzione condizionata di θ_i dati i θ_i precedenti è una mistura[24] di $\delta_{\theta_1}, \ldots, \delta_{\theta_{i-1}}$ e G_0 ed ha la seguente forma:

$$\theta_i|\theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0}\delta_{\theta_l} + \frac{\alpha_0}{i-1+\alpha_0}G_0$$

dove la sommatoria definisce la probabilità di osservare nuovamente un valore già estratto e l'ultimo addendo definisce la probabilità di osservare un nuovo valore, estratto da G_0; i valori assunti dai θ_i sono detti *atomi*. Si può interpretare il tutto come un urna con $i-1$ palline che corrispondono a $\theta_1, \ldots, \theta_{i-1}$: la probabilità di pescare un colore già visto dall'urna è proporzionale a quante volte è già stato estratto in passato, la probabilità di generare un colore mai visto è proporzionale a α_0 e il nuovo valore è estratto da G_0.

Per rendere la proprietà di clustering esplicita, si considera una rappresentazione alternativa. Siano $\phi_1, \ldots, \phi_K$ valori distinti estratti da $\theta_1, \ldots, \theta_{i-1}$ e sia m_k il numero di $\theta_{i'}$ uguali a ϕ_k per $1 \leq i' < i$, la distribuzione condizionata può essere riscritta come:

$$\theta_i|\theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha_0}\delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0}G_0$$

Il *Pólya urn scheme* può essere spiegato con la metafora del *chinese restaurant process* in cui si considera un ristorante cinese con un numero infinito di tavoli; ogni θ_i corrisponde a un cliente che entra nel ristorante mentre i valori distinti di ϕ_k corrispondono ai tavoli su cui i clienti si siedono. L'i-mo cliente può:

- sedersi al tavolo indicizzato con ϕ_k con probabilità proporzionale al numero di clienti m_k che si sono già seduti lì;
- sedersi a un nuovo tavolo con probabilità proporzionale a α_0.

Nel primo caso ci si limita a fissare $\theta_i = \phi_k$; nel secondo si incrementa K di 1, si pesca ϕ_K da G_0 e si fissa $\theta_i = \phi_K$.

A.1.3 Dirichlet Process Mixture Model

Una delle applicazioni più importanti del DP è la distribuzione a priori dei parametri di un modello mistura; siano x_i le osservazioni,

$$\theta_i|G \sim G$$
$$x_i|\theta_i \sim F(\theta_i)$$

[24] Infatti si ha $\sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0} + \frac{\alpha_0}{i-1+\alpha_0} = \frac{i-1+\alpha_0}{i-1+\alpha_0} = 1$.

dove i θ_i sono condizionatamente indipendenti dato G e le x_i lo sono dato θ_i. Se $G \sim DP(\alpha_0, g_0)$, allora il modello è detto *DP Mixture Model* e può essere rappresentato come segue:

$$\begin{aligned}
\boldsymbol{\pi}|\alpha_0 &\sim \text{GEM}(\alpha_0) \\
z_i|\boldsymbol{\pi} &\sim \boldsymbol{\pi} \\
\phi_k|G_0 &\sim G_0 \\
x_i|z_i, (\phi_k)_{k=1}^{\infty} &\sim F(\phi_{z_i})
\end{aligned}$$

Inoltre, riprendendo la notazione precedente, si ha $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$ e $\theta_i = \phi_{z_i}$.

A.2 Hierarchical Dirichlet Process

Un *processo di Dirichlet gerarchico* (*HDP*) è definito come un insieme di distribuzioni di una misura di probabilità G_j sullo spazio misurabile $(\Theta, \mathcal{B})$; il processo definisce una misura di probabilità G_j per ogni gruppo e una misura di probabilità globale G_0. G_0 è a sua volta distribuita come un DP con parametro di concentrazione γ e distribuzione di base H, le G_j sono condizionatamente indipendenti dato G_0 e distribuite come un DP con parametro di concentrazione α_0 e distribuzione di base G_0:

$$\begin{aligned}
G_0|\gamma &\sim DP(\gamma, H) \\
G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0)
\end{aligned}$$

La baseline H fornisce la distribuzione a priori per i fattori θ_{ji}. La distribuzione G_0, comune a tutti i gruppi, varia intorno a H con un ammontare di variabilità governato da γ[25] mentre le distribuzioni G_j sui fattori nei gruppi deviano dalla distribuzione comune G_0 con un ammontare di variabilità governato da α_0. Se ci si aspetta che la variabilità sia differente nei vari gruppi, è possibile utilizzare un parametro di concentrazione differente α_j per ogni gruppo.

Il *HDP* può essere facilmente esteso a più di due livelli estraendo a sua volta la baseline H da un DP, la distribuzione di base di H può essere a sua volta estratta da un DP e così via; in generale, si ottiene un albero in cui un DP è associato a ogni nodo, i figli di un nodo sono condizionatamente indipendenti dato il loro nodo padre e l'estrazione dal DP di un nodo serve come misura di base per i nodi figli. Gli atomi ad un determinato nodo sono quindi condivisi tra tutti i discendenti del nodo, fornendo quindi una nozione di cluster condivisi su più livelli.

A.2.1 The Stick-Breaking Construction

Dato che $G_0 \sim DP(\gamma, H)$, vale la seguente *stick-breaking construction*:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

dove $\phi_k \sim H$ indipendenti e $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \sim \text{GEM}(\gamma)$ sono mutuamente indipendenti. Dal momento che G_0 ha come supporto i punti $\boldsymbol{\phi} = (\phi_k)_{k=1}^{\infty}$, ogni G_j ha necessariamente supporto in

[25] Per valori alti di γ la distribuzione di G_0 tenderà a coincidere con quella di H.

questi punti[26] e quindi può essere scritto come segue:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

Sia $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty}$, i pesi $\boldsymbol{\pi}_j$ sono indipendenti dato $\boldsymbol{\beta}$ poiché i G_j sono indipendenti dato G_0; inoltre, si può dimostrare che $\boldsymbol{\pi}_j \sim DP(\alpha_0, \boldsymbol{\beta})$.

A.2.2 The Chinese Restaurant Franchise

La metafora del *chinese restaurant process* è estesa in modo da poter considerare più ristoranti che condividono lo stesso gruppo di piatti, da cui il nome *chinese restaurant franchise*. Si considera un franchise con un unico menù condiviso da tutti i ristoranti. A ogni tavolo di ogni ristorante, un piatto è ordinato dal menù dal primo primo cliente che si è seduto, il piatto è condiviso da tutti i clienti che siedono al tavolo. Più tavoli in più ristoranti possono servire lo stesso piatto.

In questo contesto, i ristoranti corrispondono ai gruppi e i clienti ai fattori θ_{ji}, il menù globale dei piatti è composto da $\phi_1, \ldots, \phi_K \overset{iid}{\sim} H$, ψ_{jt} è il piatto servito al tavolo t nel ristorante j. Siano inoltre

- n_{jtk} il numero di clienti nel ristorante j al tavolo t che mangiano il piatto k;
- $n_{jt\cdot}$ il numero di clienti nel ristorante j al tavolo t;
- $n_{j\cdot k}$ il numero di clienti nel ristorante j che mangiano il piatto k;
- m_{jk} il numero di tavoli occupati nel ristorante j che servono il piatto k;
- $m_{j\cdot}$ il numero di tavoli occupati nel ristorante j;
- $m_{\cdot k}$ il numero di tavoli occupati che servono il piatto k;
- $m_{\cdot\cdot}$ il numero di tavoli occupati.

La distribuzione condizionata di θ_{ji} dati $\theta_{j1}, \ldots, \theta_{j,i-1}$ e G_0, avendo già integrato G_j, è la seguente mistura:

$$\theta_{ji} | \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i - 1 + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i - 1 + \alpha_0} G_0$$

Si noti che la probabilità di osservare un nuovo valore –sedersi a un nuovo tavolo e scegliere un piatto selezionandolo da G_0– è proporzionale a α_0 mentre la probabilità di osservare nuovamente un valore già noto –sedersi a un tavolo e quindi scegliere il piatto di quel tavolo– è proporzionale al numero di clienti che già sono seduti in quel tavolo. Integrando G_0, è possibile ottenere la distribuzione condizionata di ψ_{jt} che è anch'essa una mistura:

$$\psi_{jt} | \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\psi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H$$

Si noti che la probabilità di osservare un nuovo valore –scegliere un piatto selezionandolo dal menù globale H– è proporzionale a γ mentre la probabilità di osservare nuovamente un valore già noto –scegliere un piatto che viene già servito in qualche tavolo– è proporzionale al numero di tavoli che

[26] Più nello specifico, se $G_j \sim DP(\alpha_0, G_0)$, allora il supporto di G_j è un sottoinsieme del supporto della distribuzione di base G_0.

servono quel piatto. Per ottenere estrazioni da θ_{ji} a partire dalle due distribuzioni condizionate si procede come segue:

1. Estraggo θ_{ji} dalla distribuzione condizionata di θ_{ji}.
2. Se è necessario estrarre un valore da G_0, estraggo ψ_{ji} dalla distribuzione condizionata di ψ_{ji} e fisso $\theta_{ji} = \psi_{ji}$.

A.2.3 Hierarchical Dirichlet Process Mixture Model

Un *HDP* può essere usato come la distribuzione a priori dei fattori per dati raggruppati. Per ogni j, siano $\theta_{j1}, \theta_{j2}, \ldots \overset{iid}{\sim} G_j$, a ogni fattore θ_{ji} corrisponde un'osservazione x_{ji}; il modello mistura è dato da

$$\begin{aligned} \theta_{ji}|G_j &\sim G_j \\ x_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

Riprendendo la *stick-breaking construction*, θ_{ji} assume il valore ϕ_k con probabilità π_{jk}, il modello può essere quindi riscritto come segue:

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim \mathrm{GEM}(\gamma) \\ \boldsymbol{\pi}_j|\alpha_0, \boldsymbol{\beta} &\sim DP(\alpha_0, \boldsymbol{\beta}) \\ z_{ji}|\boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ \phi_k|H &\sim H \\ x_{ji}|z_{ji}, (\phi_k)_{k=1}^{\infty} &\sim F(\phi_{z_{ji}}) \end{aligned}$$

Bibliografia

[BNJ03] David M. Blei, Andrew Y. Ng e Michael I. Jordan. «Latent Dirichlet Allocation». In: *J. Mach. Learn. Res.* 3.null (mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[Ros+04] Michal Rosen-Zvi et al. «The Author-Topic Model for Authors and Documents». In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. UAI '04. Banff, Canada: AUAI Press, 2004, pp. 487–494. ISBN: 0974903906.

[Gri+05] Thomas Griffiths et al. «Integrating Topics and Syntax». In: *Advances in Neural Information Processing Systems*. A cura di L. Saul, Y. Weiss e L. Bottou. Vol. 17. MIT Press, 2005. URL: https://proceedings.neurips.cc/paper/2004/file/ef0917ea498b1665ad6c701057155abe-Paper.pdf.

[BL06] David M. Blei e John D. Lafferty. «Dynamic Topic Models». In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 113–120. ISBN: 1595933832. DOI: 10.1145/1143844.1143859.

[Teh+06] Yee Whye Teh et al. «Hierarchical Dirichlet Processes». In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581. DOI: 10.1198/016214506000000302.

[Wal06] Hanna M. Wallach. «Topic Modeling: Beyond Bag-of-Words». In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 977–984. ISBN: 1595933832. DOI: 10.1145/1143844.1143967.

[WM06] Xuerui Wang e Andrew McCallum. «Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends». In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 424–433. ISBN: 1595933395. DOI: 10.1145/1150402.1150450.

[BL07] David M. Blei e John D. Lafferty. «A correlated topic model of Science». In: *The Annals of Applied Statistics* 1.1 (2007), pp. 17–35. DOI: 10.1214/07-AOAS114.

[Nal+07] Ramesh M. Nallapati et al. «Multiscale Topic Tomography». In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: Association for Computing Machinery, 2007, pp. 520–529. ISBN: 9781595936097. DOI: 10.1145/1281192.1281249.

[SG07] M. Steyvers e T. Griffiths. «Latent Semantic Analysis: A Road to Meaning». In: a cura di T. Landauer, S. Dennis McNamara e W. Kintsch. Laurence Erlbaum, 2007. Cap. Probabilistic topic models.

[WBH08] Chong Wang, David Blei e David Heckerman. «Continuous Time Dynamic Topic Models». In: *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. UAI'08. Helsinki, Finland: AUAI Press, 2008, pp. 579–586. ISBN: 0974903949.

[Ram+09] Daniel Ramage et al. «Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora». In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 2009, pp. 248–256. URL: https://aclanthology.org/D09-1026.

[BCD10] David Blei, Lawrence Carin e David Dunson. «Probabilistic Topic Models». In: *IEEE Signal Processing Magazine* 27.6 (2010), pp. 55–65. DOI: 10.1109/MSP.2010.938079.

[Zha+10] Jianwen Zhang et al. «Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-Varying Corpora». In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. Washington, DC, USA: Association for Computing Machinery, 2010, pp. 1079–1088. ISBN: 9781450300551. DOI: 10.1145/1835804.1835940.

[Tsa11] Flora S. Tsai. «A Tag-Topic Model for Blog Mining». In: *Expert Syst. Appl.* 38.5 (mag. 2011), pp. 5330–5335. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2010.10.025.

[Zha+11] Wayne Xin Zhao et al. «Comparing Twitter and Traditional Media Using Topic Models». In: *Advances in Information Retrieval*. A cura di Paul Clough et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 338–349. ISBN: 978-3-642-20161-5.

[Ble12] David M. Blei. «Probabilistic Topic Models». In: *Commun. ACM* 55.4 (apr. 2012), pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826.

[Dub+13] Avinava Dubey et al. «A nonparametric mixture model for topic modeling over time». In: *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)*. 2013, pp. 530–538. DOI: 10.1137/1.9781611972832.59.

[Ma+13] Zhiqiang Ma et al. «Tag-Latent Dirichlet Allocation: Understanding Hashtags and Their Relationships». In: *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*. WI-IAT '13. USA: IEEE Computer Society, 2013, pp. 260–267. ISBN: 9780769551456. DOI: 10.1109/WI-IAT.2013.38.

[Lin+14] Tianyi Lin et al. «The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text». In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. Seoul, Korea: Association for Computing Machinery, 2014, pp. 539–550. ISBN: 9781450327442. DOI: 10.1145/2566486.2567980.

[Li+16] S. Li et al. «Correlated tag learning in topic model». In: *32nd Conference on Uncertainty in Artificial Intelligence 2016, UAI 2016* (2016), pp. 457–466.

[Rud+16] Maja Rudolph, Francisco J. R. Ruiz et al. «Exponential Family Embeddings». In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 478–486. ISBN: 9781510838819.

[Zha+16] Feng Zhao et al. «A Personalized Hashtag Recommendation Approach Using LDA-Based Topic Model in Microblog Environment». In: *Future Gener. Comput. Syst.* 65.C (dic. 2016), pp. 196–206. ISSN: 0167-739X. DOI: 10.1016/j.future.2015.10.012.

[BHM17] Jordan Boyd-Graber, Yuening Hu e David Mimno. «Applications of Topic Models». In: *Foundations and Trends® in Information Retrieval* 11.2-3 (2017), pp. 143–296. ISSN: 1554-0669. DOI: 10.1561/1500000030.

[Rud+17] Maja Rudolph, Francisco Ruiz et al. «Structured Embedding Models for Grouped Data». In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 250–260. ISBN: 9781510860964.

[RB18] Maja Rudolph e David Blei. «Dynamic Embeddings for Language Evolution». In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1003–1011. ISBN: 9781450356398. DOI: 10.1145/3178876.3185999.

[Jel+19] Hamed Jelodar et al. «Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey». In: *Multimedia Tools Appl.* 78.11 (giu. 2019), pp. 15169–15211. ISSN: 1380-7501. DOI: 10.1007/s11042-018-6894-4.

[DRB20] Adji B. Dieng, Francisco J. R. Ruiz e David M. Blei. «Topic Modeling in Embedding Spaces». In: *Transactions of the Association for Computational Linguistics* 8 (lug. 2020), pp. 439–453. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00325.