Detecting and Characterizing Events: Appendices

Anonymous EMNLP submission

A Inference

In this appendix, we describe the details of the inference algorithm for Capsule. Source code for this algorithm is available at is available at https://github.com/?????/capsule.

Conditional on a collection of observed documents, our goal is to estimate the posterior values of the hidden parameters, according to the Capsule model. Recall that our data is observed as word counts $w_{d,v}$ for document d and vocabulary term v, with corresponding author and time interval information for each document—a_d and i_d, respectively. The latent parameters of the model include general topics β, entity topics η, and interval topics π, as well as document-specific strengths in each of these spaces: document relevancy to general topics θ, entity topics ζ, and interval topics ϵ. The latent parameters also include entity concerns with general topics ϕ and with entity-specific topics ξ, and overall event strength ψ. See Figure 3 for the full generative model.

As for many Bayesian models, the exact posterior for Capsule is not tractable to compute; we must instead approximate it. Thus, we develop an approximate inference algorithm for Capsule based on variational methods (Jordan et al., 1999; Wainwright and Jordan, 2008).

Variational inference approaches the problem of posterior inference by minimizing the KL divergence from an approximating distribution q to the true posterior p. This is equivalent to maximizing the ELBO,

$$\mathcal{L}(q) = \mathbb{E}_q\big[\log p(w, \psi, \pi, \phi, \beta, \xi, \eta, \theta, \epsilon, \zeta) - \log q(\psi, \pi, \phi, \beta, \xi, \eta, \theta, \epsilon, \zeta)\big]. \quad (6)$$

We define the approximating distribution q using the mean field assumption:

$$q(\psi, \pi, \phi, \beta, \xi, \eta, \theta, \epsilon, \zeta) = \prod_{d=1}^{D}\left[q(\zeta_d \mid \lambda_d)\prod_{k=1}^{K} q(\theta_{d,k} \mid \lambda^{\theta}_{d,k})\prod_{t=1}^{T} q(\epsilon_{d,t} \mid \lambda^{\epsilon}_{d,t})\right] \prod_{t=1}^{T}\left[q(\pi_t \mid \lambda^{\pi}_t) q(\psi_t \mid \lambda^{\psi}_t)\right] \prod_{n=1}^{N}\left[q(\xi_n \mid \lambda^{\xi}_n)\, q(\eta_n \mid \lambda^{\eta}_n)\right] \prod_{k=1}^{K}\left[q(\beta_k \mid \lambda^{\beta}_k)\prod_{n=1}^{N} q(\phi_{n,k} \mid \lambda^{\phi}_{n,k})\right] \quad (7)$$

The variational distributions for the topics $q(\pi)$, $q(\beta)$ and $q(\eta)$ are all Dirichlet-distributed with free variational parameters λ^{π}, λ^{β}, and λ^{η} respectively. Similarly, the variational distributions $q(\psi)$, $q(\phi)$, $q(\xi)$, $q(\theta)$, $q(\epsilon)$, and $q(\zeta)$ are all gamma-distributed with corresponding free variational parameters λ^{ψ}, λ^{ϕ}, λ^{ξ}, λ^{θ}, λ^{ϵ} and λ^{ζ}. For these gamma-distributed variables, each free parameter λ has two components: shape s and rate r.

The expectations under q, which are needed to maximize the ELBO, have closed form analytic updates–we update each parameter in turn, following standard coordinate ascent variational inference techniques, as the Capsule model is specified with the required conjugate relationships that make this approach possible (Ghahramani and Beal, 2001).

To obtain simple updates, we first rely on auxiliary latent variables z. These variables, when marginalized out, leave the original model intact. The Poisson

distribution has an additive property; specifically if $w \sim \text{Poisson}(a+b)$, then $w = z_1 + z_2$, where $z_1 \sim \text{Poisson}(z_1)$ and $z_2 \sim \text{Poisson}(z_2)$. We apply this decomposition to the word count rate in Equation (1) and define Poisson variables for each component of the word count:

$$z^{\mathcal{K}}_{d,v,k} \sim \text{Poisson}(\theta_{d,k}\beta_{k,v}),$$

$$z^{\mathcal{E}}_{d,v} \sim \text{Poisson}(\zeta_d \eta_{a_d,v}),$$

$$z^{\mathcal{T}}_{d,v,t} \sim \text{Poisson}\left(f(i_d,t)\epsilon_{d,t}\pi_{t,v}\right).$$

The $\mathcal{K}$, $\mathcal{E}$, and $\mathcal{T}$ superscripts indicate the contributions from general, entity, and event topics, respectively. Given these variables, the total word count is deterministic:

$$w_{d,v} = \sum_{k=1}^{K} z^{\mathcal{K}}_{d,v,k} + z^{\mathcal{E}}_{d} + \sum_{t=1}^{T} z^{\mathcal{T}}_{d,v,t}.$$

Coordinate-ascent variational inference is derived from complete conditionals, i.e., the conditional distributions of each variable given the other variables and observations. These conditionals define both the form of each variational factor and their updates. The following are the complete conditional for each of the gamma- and Dirchlet-distributed latent parameters. The notation $D(n)$ is used for the set of documents sent by entity n; $D(t)$ is the set of documents sent impacted by events at time t (e.g., all documents after the event in the case of exponential decay).

$$\pi_t \mid \mathbf{W}, \psi, \phi, \xi, \beta, \eta, \theta, \epsilon, \zeta, z \sim \text{Dirichlet}_V\left(\alpha_\pi + \sum_{d=1}^{D} \langle z^{\mathcal{T}}_{d,1,t}, \cdots, z^{\mathcal{T}}_{d,V,t}\rangle\right) \quad (8)$$

$$\eta_n \mid \mathbf{W}, \psi, \phi, \xi, \beta, \pi, \theta, \epsilon, \zeta, z \sim \text{Dirichlet}_V\left(\alpha_\eta + \sum_{d\in D(n)} \langle z^{\mathcal{E}}_{d,v}, \cdots, z^{\mathcal{E}}_{d,v}\rangle\right) \quad (9)$$

$$\beta_k \mid \mathbf{W}, \psi, \phi, \xi, \pi, \eta, \theta, \epsilon, \zeta, z \sim \text{Dirichlet}_V\left(\alpha_\beta + \sum_{d=1}^{D} \langle z^{\mathcal{K}}_{d,1,k}, \cdots, z^{\mathcal{K}}_{d,V,k}\rangle\right) \quad (10)$$

$$\psi_t \mid \mathbf{W}, \phi, \xi, \beta, \pi, \eta, \theta, \epsilon, \zeta, z \sim \text{Gamma}\left(s_\psi + |D(t)|s_\epsilon, r_\psi + \sum_{d\in D(t)} \epsilon_{d,t}\right) \quad (11)$$

$$\xi_n \mid \mathbf{W}, \psi, \phi, \beta, \pi, \eta, \theta, \epsilon, \zeta, z \sim \text{Gamma}\left(s_\xi + |D(n)|s_\zeta, r_\xi + \sum_{d\in D(n)} \zeta_d\right) \quad (12)$$

$$\phi_{n,k} \mid \mathbf{W}, \psi, \xi, \beta, \pi, \eta, \theta, \epsilon, \zeta, z \sim \text{Gamma}\left(s_\phi + |D(n)|s_\theta, r_\phi + \sum_{d\in D(n)} \theta_{d,k}\right) \quad (13)$$

$$\theta_{d,k} \mid \mathbf{W}, \psi, \phi, \xi, \beta, \pi, \eta, \epsilon, \zeta, z \sim \text{Gamma}\left(s_\theta + \sum_{v=1}^{V} z^{\mathcal{K}}_{d,v,k}, \phi_{a_d,k} + \sum_{v=1}^{V} \beta_{k,v}\right) \quad (14)$$

$$\epsilon_{d,t} \mid \mathbf{W}, \psi, \phi, \xi, \beta, \pi, \eta, \theta, \zeta, z \sim \text{Gamma}\left(s_\epsilon + \sum_{v=1}^{V} z^{\mathcal{T}}_{d,v,t}, \psi_t + f(i_d,t)\sum_{v=1}^{V} \pi_{t,v}\right) \quad (15)$$

$$\zeta_d \mid \mathbf{W}, \psi, \phi, \xi, \beta, \pi, \eta, \theta, \epsilon, z \sim \text{Gamma}\left(s_\zeta + \sum_{v=1}^{V} z^{\mathcal{E}}_{d,v}, \xi_{a_d} + \sum_{v=1}^{V} \eta_{a_d,v}\right) \quad (16)$$

The complete conditional for the auxiliary variables has the form $z_{d,v} \mid \psi, \phi, \xi, \beta, \pi, \eta, \theta, \epsilon, \zeta \sim \text{Mult}(w_{d,v}, \omega_{d,v})$, where

$$\omega_{d,v} \propto \langle \theta_{d,1}\beta_{1,v}, \cdots, \theta_{d,K}\beta_{K,v}, \zeta_d\eta_{a_d,v}, f(i_d,1)\epsilon_{d,1}\pi_{1,v}, \cdots, f(i_d,T)\epsilon_{d,T}\pi_{T,v}\rangle. \quad (17)$$

Intuitively, these variables allocate the data to one of the entity concerns or events, and thus can be used to explore the data.

Algorithm 1: Variational Inference for Capsule

Input: word counts w
Output: approximate posterior of latent parameters in terms of variational parameters $\lambda = \{\lambda^{\psi}, \lambda^{\pi}, \lambda^{\xi}, \lambda^{\eta}, \lambda^{\phi}, \lambda^{\beta}, \lambda^{\theta}, \lambda^{\zeta}, \lambda^{\epsilon}\}$
Initialize $\mathbb{E}[\beta]$ to slightly random around uniform
Initialize $\mathbb{E}$[all other parameters] to uniform
for *iteration* $m = 1 : M$ **do**
  set $\lambda^{\psi}, \lambda^{\pi}, \lambda^{\xi}, \lambda^{\eta}, \lambda^{\phi}, \lambda^{\beta}, \lambda^{\theta}, \lambda^{\zeta}, \lambda^{\epsilon}$ to respective priors, excluding $\lambda^{\theta,rate}$, $\lambda^{\zeta,rate}$, and $\lambda^{\epsilon,rate}$, which are set to 0
  update $\lambda^{\theta,rate} \mathrel{+}= \sum_V \mathbb{E}[\beta_v]$
  for *each document* $d = 1 : D$ **do**
    for *each term* $v \in V(d)$[1] **do**
      set $(K+T+1)$-vector $\omega_{d,v}$ as shown in eq. (17), using $\mathbb{E}$ of parameters
      set $(K+T)$-vector $\mathbb{E}[z_{d,v}] = w_{d,v} * \omega_{d,v}$
      update $\lambda_d^{\theta,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ [eq. (14)]
      update $\lambda_d^{\epsilon,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ [eq. (15)]
      update $\lambda_d^{\zeta,shape} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{E}}]$ [eq. (16)]
      update $\lambda_v^{\beta} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{K}}]$ [eq. (10)]
      update $\lambda_v^{\pi} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{T}}]$ [eq. (8)]
      update $\lambda_v^{\eta} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{E}}]$ [eq. (9)]
    end
    set $\lambda_d^{\theta,rate} = \mathbb{E}[\phi_{a_d}] + \sum_v \mathbb{E}[\beta]$ [eq. (14)]
    set $\lambda_d^{\epsilon,rate} = \mathbb{E}[\psi] + f \sum_v \mathbb{E}[\pi]$ [eq. (15)]
    set $\lambda_d^{\zeta,rate} = \mathbb{E}[\xi_{a_d}] + \sum_v \mathbb{E}[\eta]$ [eq. (16)]
    set $\mathbb{E}[\theta_d] = \lambda_d^{\theta,shape} / \lambda_d^{\theta,rate}$
    set $\mathbb{E}[\epsilon_d] = \lambda_d^{\epsilon,shape} / \lambda_d^{\epsilon,rate}$
    set $\mathbb{E}[\zeta_d] = \lambda_d^{\zeta,shape} / \lambda_d^{\zeta,rate}$
    update $\lambda_{a_d}^{\phi,shape} \mathrel{+}= s_{\theta}$ [eq. (13)]
    update $\lambda_t^{\psi,shape} \mathrel{+}= s_{\epsilon} \forall t : f(i_d, t) \neq 0$ [eq. (11)]
    update $\lambda_{a_d}^{\xi,shape} \mathrel{+}= s_{\eta}$ [eq. (12)]
    update $\lambda_{a_d}^{\phi,rate} \mathrel{+}= \theta_d$ [eq. (13)]
    update $\lambda^{\psi,rate} \mathrel{+}= \epsilon_d$ [eq. (11)]
    update $\lambda_{a_d}^{\xi,rate} \mathrel{+}= \zeta_d$ [eq. (12)]
  end
  set $\mathbb{E}[\phi] = \lambda^{\phi,shape} / \lambda^{\phi,rate}$
  set $\mathbb{E}[\beta_k] = \lambda^{\beta_{k,v}} / \sum_v \lambda^{\beta_k} \forall k$
  set $\mathbb{E}[\xi] = \lambda^{\xi,shape} / \lambda^{\xi,rate}$
  set $\mathbb{E}[\eta_n] = \lambda^{\eta_{n,v}} / \sum_v \lambda^{\eta_n} \forall n$
  set $\mathbb{E}[\psi] = \lambda^{\psi,shape} / \lambda^{\psi,rate}$
  set $\mathbb{E}[\pi_t] = \lambda^{\pi_{t,v}} / \sum_v \lambda^{\pi_t} \forall t$
end
return λ

Given these conditionals, the algorithm sets each parameter to the expected conditional parameter under the variational distribution. The mean field assumption guarantees that this expectation will not involve the parameter being updated. Algorithm 1 shows our variational inference algorithm.

B Additional Results

In this appendix, we present non-crucial experimental results for Capsule.

Events detected by Capsule. In addition to the real-world events discussed in Section 3, many other events were detected by Capsule. For example, in 1973, the U.S. presented all nations with samples of rock obtained from the moon on the Apollo 17 mission. Top cables under ϵ for this week include messages about receiving the sample shipments at the various embassies, such as a message sent by the Cairo embassy on July 2, 1973:

> Lunar sample recieved June 30. Will advise Department of our plans for presentation as soon as they are firm.

and a similar message from the Kuala Lumpur Embassy the same day:

> Lunar sample and accompanying materials unreceived. Advise shipping data.

We also observe a cable sent by New Delhi the following day:

> 1. As Department aware, lunar sample for India was presented by Apollo 17 austronauts to Lok Sabha (Lower House of Parliment) Speaker G.S. Dhillon on Jun 19. Captain Cernan's remarks on that occasion on which he was escorted by the Charge', were similar to those suggested Ref B.
>
> 2. Embassy has received lunar sample for Bhutan (Ref A) which will be presented to Bhutanese on some propitious future date. In

[1] $V(d)$ is the set of vocabulary indices for the collection of words in document d. We could also iterate over all V, but as zero word counts give $\mathbb{E}[z_{d,v}] = 0 \; \forall v \notin V(d)$, the two are equivalent.

> making presentation, remarks suggested Ref B will be drawn upon as appropriate.

Another peak in Figure 1 occurs the week of April 17, 1978 surrounding a UN special session on disarmament; the top three words under event its description π are *SSOD* (acronym for "special session on disarmament", *disarmament*, and *ICS* (likely an acronym for "incident command system"); Table 6 shows the top cables for this time interval, sorted by event relevancy ϵ. Most of these cables concern attendance, such as the first cable, which is from Bangkok:

> Thai Ministry of Foreign Affairs official in International Organizations Department told EmbOff Apr 19 that his Dept had recommended that Foriegn Minister Uppadit attend SSOD. Uppadit, however, had just arrived back from trip to Asean countries and had not yet considered composition of delegation. MFA official expected the Foreign Minister would not decided on his attendance until early May. Will advise.

Unfortunately, many interesting cables are withdrawn, meaning the main body of the message is unavailable. An example of this is the cable sent on April 20, 1978 from the State Department to London with the subject *SSOD: security assurances for non-nuclear states.* The following description is provided by NARA for withdrawn cables such as these:

> *The Department of State created withdrawal cards for telegrams exempt from disclosure for reasons of national security, privacy, and other statutory concerns. The agency's withdrawal cards serve as placeholders for the records exempt from disclosure. NARA created withdrawal cards for an additional set of telegrams that require review under the Freedom of Information Act. The data elements potentially available for each withdrawal card include: concepts; date; document number; from; subject; traffic analysis by geography and subject; to; and microfilm roll number.*

The return of the crown of St. Stephen to Hungary was a contentious event, with many individuals sending in their opinions on it; Table 7 shows top cables for this event.

In early October of 1977, the International Whaling commission (IWC) banned killing bowhead whales. Capsule detects this surge in discussion and the top cables recovered by Capsule indicate that the U.S. State department objected to this ban,[2] but that many individuals disagreed, resulting in the majority of the top cables having subjects along the lines of *Requests State Dept not file objection to zero quota for bowhead whales*.

Similar to the sequence of Sinai events, another sequence of events occurs surrounding opium production; in March 1974, Capsule detects that Turkey plans to lift a ban on growing opium poppy. Two months later, the model detects another event when the U.S. makes a policy statement on the domestic production of opium poppy. For the first event, the top three cables for this interval have subjects *Turkey to resume cultivation of the opium poppy*, *Cultivation of opium poppy in Turkey*, and *Ban on opium cultivation in Turkey*. Another cable further down the list has the subject, *US funds to Turkey to halt opium production*. On March 12, 1975, the London embassy reports the the State Department:

> 1. Pursuant to RefTel, we spoke to W. N. Hillier-Fry, head of Middle East and Mediterranean Department of Overseas Development Administration, who will head UK delegation to meeting of OECD consortium for Turkey. Hillier-Fry said that, as in last meeting, he can support USG initiatives with regard to Turkish opium production, but he emphasized that he will have to take a very cautious approach and will pick his words carefully. He added that he will be in touch with USG representatives at the meeting.
>
> 2. As FCO department head in separate conversation March 11, UK representative will have to be careful in whatever he says not to imply any increase in British aid to Turkey, given straitened financial circumstances of

[2] Their objections were based on Alaskan Natives rely on these whales for sustenance.

ϵ	date	entity	subject
0.084	1978-04-19	Bangkok	UN Special Session on Disarmament: high level participation: Thailand
0.074	1978-04-19	Valletta	UN Special Session on Disarmament, May 23 - June 28: high level participation
0.073	1978-04-20	Bern	UN Special Session on Disarmament, May 23-June 28: high- level participation
0.073	1978-04-23	State	High level attendance at Special Session on Disarmament (SSOD)
0.073	1978-04-21	Lagos	UN Special Session on Disarmament, May 23-June 28: high level participation
0.069	1978-04-19	Madrid	UN Special Session on Disarmament May 23 - June 28: high-level participation
0.067	1978-04-19	Seoul	UN Special Session on Disarmament
0.067	1978-04-19	Tehran	Iranian participation at SSOD May 23-June 28
0.065	1978-04-19	Niamey	UN Special Session on Disarmament, May 23-June 26 high level participation
0.061	1978-04-19	Berlin	UN Special Session on Disarmament, May 23 - June 28; high-level participation
0.061	1978-04-18	Sofia	official-informal
0.060	1978-04-19	Abidjan	UN Special Session on Disarmament: Ivorian participation
0.056	1978-04-19	Athens	UN Special Session on Disarmament, May 23-June 28: high level participation
0.054	1978-04-20	State	SSOD: security assurances for non-nuclear states
0.053	1978-04-19	Kuwait	UN Special Session on Disarmament (SSOD), May 23-June 28: high level ...
0.051	1978-04-19	Vienna	UN Special Session on Disarmament, May 23-June 28: high level part...
0.050	1978-04-20	Bogota	Meeting delegations: UN Special Session on Disarmament

Table 6: Top documents for the time interval of week April 17, 1978, in preparation for the a UN Special Session on Disarmament.

> hmg at present. Goodison also confirmed letter from him has gone to British ambassador in Ankara along lines indicated London's 2980.

Two days later, Kissinger sends a message to Ankara with the subject *Opium ban: attitudes within GOT*:

> In conversation with Sisco reported RefTel Esenbel referred to division within got on handling of opium problems. At one point he identified Orhan Eyuboglu as leader of hard liners who wished immediate resumption of poppy cultivation. By implication he identified Fonmin Gunes and himself as members of those seeking cooperate with united states, e.g. Gunes "very pleased" with decision not to plant in April.

Eight weeks later, we see the second event with many messages on individuals writing on the subject of the *US policy statement on domestic production of opium poppy*. Again, Capsule cannot capture every aspect of larger sequences of events, but it can provide insight into key moments, as it does here.

Topics Found by Capsule. Table 8 shows a selection of general topics, including those shown in Tables 1 and 4.

Table 9 shows a selection of entity topics, including those shown in Tables 5 and 4.

Table 10 shows a selection of event topics.

arXiv event detection. Figure 1 shows a variety of events that Capsule detected from the National Archive diplomatic cables data, many of which were discussed in Section 3 and above. To provide a negative example (Capsule not detecting events), we consider a collection of arXiv abstracts posted from 1995 through part of 2016. As these scientific paper pre-print abstracts should not have event resolution on the order of weeks, we do not anticipate finding meaningful events in this corpus.

Figure 6 shows that Capsule does not detect events on this data, as anticipated. Early in the time range, when there is little data, the measure of "eventness" fluctuates wildly. The measure does produce a few peaks, but these are a result of over-fitting rather than the true detection of real-world events–the top abstracts at these time points do not a reveal a consistent theme. This process of checking the top documents for each of the rare peaks is quick to perform.

Model Sensitivity. Using simulated data, we assessed the sensitivity of our model to different decay functions f and decay durations τ. We simulated data following the description in Section 3. We considered an exponential decay function, shown in

ϵ	date	entity	subject
0.088	1978-01-05	Sarasin, Ronald A	Return of the crown of St Stephen to Hungary
0.088	1978-01-05	Cotter, William R	Return of the crown of St Stephen to Hungary
0.080	1978-01-05	Cranston, Alan	Return of crown of St Stephen to Hungary
0.078	1978-01-06	Schweiker, Richard S	Return of crown of St Stephen to Hungary
0.076	1978-01-06	Church, Frank	Return of crown of Stepeht to Hungary
0.076	1978-01-05	Wright, Jim	Return of crown of St Stephen to Hungary
0.076	1978-01-05	Wright, Jim	Against return of crown of St Stephen to Hungary
0.068	1978-01-03	Tarnoff, Peter	Return of the crown of St Stephen to Hungary
0.066	1978-01-03	Tarnoff, Peter	Return of crown of St Stephen to Hungary
0.056	1978-01-03	Duncan, John J	Info on the crown of St Stephen of Hungary
0.053	1978-01-06	Church, Frank	Return of crown of St Stephen to Hungarian govt
0.052	1978-01-06	Nelson, Gaylord	Cost of Emperor Bokassa's coronation in Central African Empire
0.051	1978-01-03	Beckel, Robert G	return of crown of St Stephen to Hungary
0.049	1978-01-05	Cranston, Alan	Concern regarding the crown of St Stephen of Hungary
0.046	1978-01-03	Chiles, Lawton	Against returning crown of St Stephen to Hungarian government
0.046	1978-01-05	Secretary Paris	The crown: Mrs. Vance's schedule
0.043	1978-01-05	Beckel, Robert G	RE constituents concern over return of crown of St. Stephen

Table 7: Top documents for the time interval of week of January 2, 1978, when Pres. Carter decided to return teh crown of St. Stephen to Hungary.

top terms
plan, visit, arrival, itinerary, visitor
outlook, review, hire, personnel, invite, prepare
arrest, incident, security, family, guard, death, jail
locate, home, son, death, please, contact, father
request, refugee, response, service, sale, asylum
market, report, commercial, food, import, commerce
fear, leadership, back, arm, role, threaten
hotel, travel, reservation, visit, arrange, schedule
exchange, student, rate, assume, program, cultural
copy, publication, panama, brochure, material, order
registry, pouch, number, invoice, item, classify
extension, provision, decide, case, decision, effect
right, cause, history, solution, improve, relation
nation, peace, soviet, crisis, strengthen, victory
israel, israeli, middle, concern, charge, negotiation
fund, management, project, overseas, committee
large, industry, sell, sale, supplier, limit, firm
concern, right, belief, point, fact, allege, explain
support, election, success, war, leader, demonstrate
case, visa, arrive, eta, consul, embassy, travel, reftel
science, advisory, concern, study, request, follow

Table 8: Top vocabulary terms for a selection of general topics, one per row, according to topic distributions β_k.

Equation (3), as well as linear decay,

$$f(i_d, t) = \begin{cases} 1 - \frac{i_d - t}{\tau + 1}, & \text{if } t \leq i_d < t + \tau \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

and a step function,

$$f(i_d, t) = \begin{cases} 1, & \text{if } t \leq i_d \leq t + \tau \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

We simulated data with $\tau = 3$ and simulated ten data sets for each of the three functions f.

In fitting the models, we also considered all three functions f, but we also varied the decay duration τ from 1 to 5. Figure 7 shows the results of these experiments, using both event detection and document recovery metrics discussed in Section 3.

For event detection, the various decay functions perform roughly comparably. For document recovery, the exponential function has mild advantages on all data types. In exploring results on the real-world cable data, we found that the exponential decay provided the mos interpretable results.

References

Zoubin Ghahramani and Matthew J Beal. 2001. Propagation algorithms for variational bayesian learning.

entity	top terms
Algiers	algerian, algeria, say, one, embassy, do
Amman	jordanian, visit, make, say, time, do, meet
Bangkok	bangkok, thailand, thai, refugee, follow
Barcelona	spain, spanish, lane, repatriation
Beirut	beirut, lebanon, lebanese, say, report
Brussels	belgian, brussels, belgium, meet, embassy
Budapest	hungarian, embassy, hungary, visit, mudd
Cape Town	cape, town, africa, say, african, make
Caracas	venezuelan, venezuela, caracas, embassy
Casablanca	casablanca, morocco, moroccan, request
Dublin	irish, goi, make, say, government, time,
Dusseldorf	opportunity, license, german, reverse
Florence	consulate, italian, party, italy, communist
Fukuoka	japan, consulate, tokyo, billion, resale
Geneva	meet, geneva, follow, make, state, take,
Guadalajara	mexico, mexican, prisoner, congen, review
Guayaquil	ecuador, ecuadorean, general, congen, one
Helsinki	finnish, helsinki, embassy, meet, make, one
Hermosillo	mexico, prisoner, attorney, aircraft, release
Jerusalem	jerusalem, israeli, bank, report, say
Johannesburg	africa, trade, african, firm, black, report
Kampala	ugandan, nairobi, african, imperialist
Kathmandu	cargo, embassy, nepalese, make, heck, report
Kingston	jamaican, minister, government, state, meet
Lagos	lagos, nigerian, nigeria, state, embassy, make
London	london, meet, british, make, time, say, follow
Mexico	mexico, mexican, embassy, gom, make, state, meet
Moscow	soviet, moscow, embasy, ussr, meet
Ndjamean	chadian, chad, lagos, drought, austerity
New Delhi	india, indian, delhi, goi, make, say, state
Oslo	norwegian, norway, embassy, meet, minister, say
Panama	panama, panamanian, embassy, canal, gop, meet
Paris	paris, french, rush, meet, france, make, say
Rome	italian, rome, meet, italy, make, follow, state
Seoul	korean, korea, seoul, rok, rokg, embassy, make
State	request, follow, embassy, meet, make
Stockholm	swedish, sweden, trade, meet, embassy
Vancouver	government, canada, canadian, columbia, say
Vienna	vienna, austrian, meet, make, follow, state
Zagreb	yugoslavia, yugoslav, general, note, croatian
Zurich	swiss, congen, bern, bank, franc, dollar, sec

Table 9: Top vocabulary terms for a selection of entities according to entity-exclusive topics η_n.

week	event	top terms
1975-12-01	Indonesian invasion of East Timor	ciec, timor, decolonization, gsp
1975-04-21	evacuation of Saigon	vietnam, evacuation, evacuate, missionary
1973-07-02	Apollo 17 lunar gifts	apollo, icc, sample, decon
1976-06-28	US bicentennial and Operation Entebbe	bicentennial, hijack, mercenary, aub
1978-04-17	UN special session on disarmament	ssod, disarmament, ica, intl
1977-10-03	IWC bans hunting of bowhead whales	whale, quota, zero, iwc
1974-03-11	Turkey plans to life ban on opium poppy	opium, poppy, omb, turkey
1974-05-06	US statement on opium poppy	rush, poppy, splex, opium
1975-09-08	Sinai Interim agreement	sinai, marianas, sept, iatf
1975-10-13	hiring for Sinai peacekeeping force	sinai, constituent, volunteer, employment
1976-09-06	death of Mao Tse-tung	sept, intelsat, tung, amspec
1978-01-02	crown of St. Stephen returned to Hungary	crown, hungary, jan, dpob

Table 10: Top vocabulary terms for a selection of time intervals according to event topics π_t.

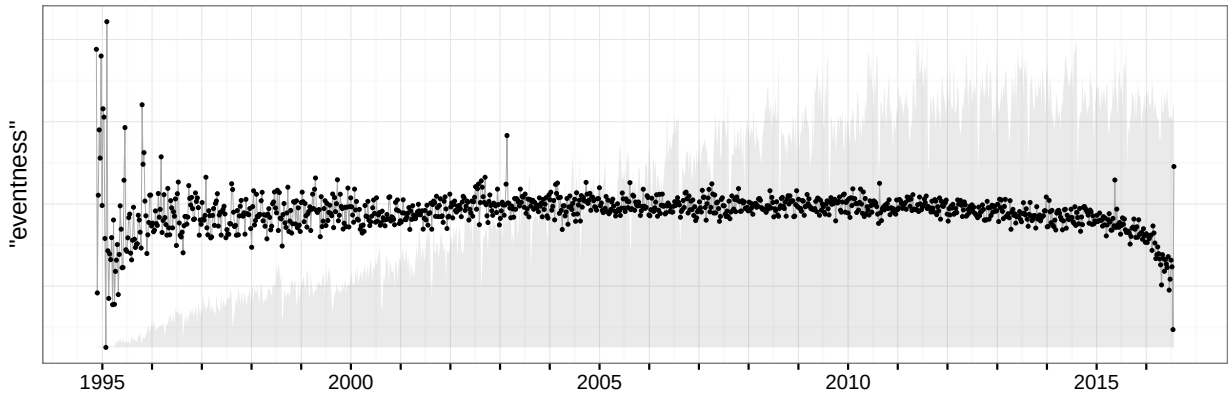

Figure 6: Measure of "eventness" over arXiv content, Equation (2). Grey background indicates the number of abstracts submitted over time.

Advances in neural information processing systems, pages 507–513.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January.

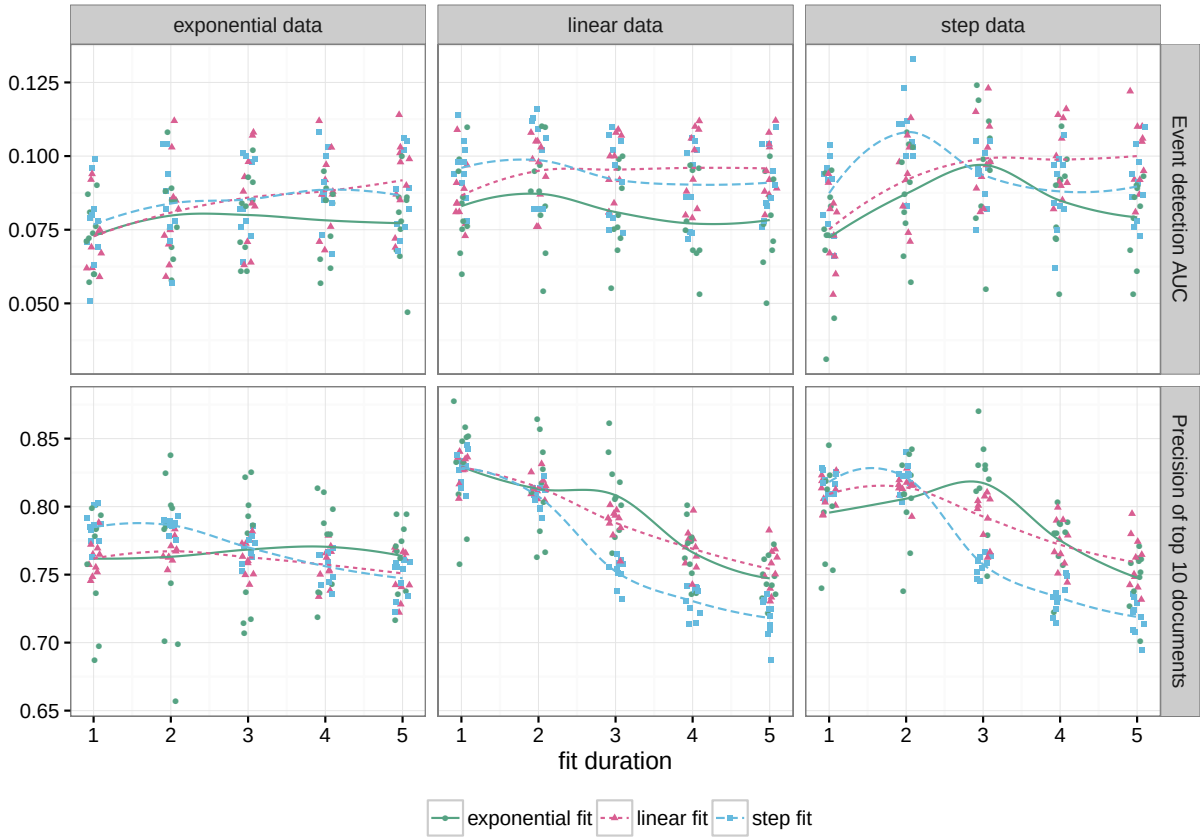

Figure 7: Assessment of model parameter sensitivity on simulated data–exponential decay tends to perform the best for document recovery, but by a small margin.