Detecting and Characterizing Events: Appendices

Allison J.B. Chaney
Princeton University
achaney@cs.princeton.edu

Hanna Wallach
Microsoft Research
wallach@microsoft.com

David M. Blei
Columbia University
david.blei@columbia.edu

Matthew Connelly
Columbia University
mjc96@columbia.edu

A Inference

In this appendix, we describe the details of the inference algorithm for 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 γ, 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 4 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,

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

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

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

The variational distributions for the topics $q(\gamma)$, $q(\beta)$ and $q(\eta)$ are all Dirichlet-distributed with free variational parameters λ^{γ}, 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 property to the word count rate in Equation (1) and define Poisson variables for each component of the word count:

$$z^{\mathcal{K}}_{d,v,k} \sim \text{Poisson}(\theta_{d,k}\beta_{k,v}),$$
$$z^{\mathcal{E}}_{d,v} \sim \text{Poisson}(\zeta_d \eta_{a_d,v}),$$
$$z^{\mathcal{T}}_{d,v,t} \sim \text{Poisson}\left(f(i_d,t)\epsilon_{d,t}\gamma_{t,v}\right).$$

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).

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

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

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

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

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

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

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

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

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

The complete conditional for the auxiliary variables has the form

$$z_{d,v} \mid \psi, \phi, \xi, \beta, \gamma, \eta, \theta, \epsilon, \zeta \sim \text{Mult}(w_{d,v}, \omega_{d,v}),$$

where

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

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

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.

This algorithm uses the notation $\boldsymbol{\lambda}$ to refer to the set of variational parameters,

$$\boldsymbol{\lambda} = \{\lambda^{\gamma}, \lambda^{\eta}, \lambda^{\beta}, \lambda^{\psi}, \lambda^{\zeta}, \lambda^{\phi}, \lambda^{\theta}, \lambda^{\epsilon}, \lambda^{\xi}\}.$$

The notation $V(d)$ 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.

This algorithm produces a fitted variational distribution which can then be used as a proxy for the true posterior, allowing us to explore a collection of documents with Capsule. Source code is available at https://github.com/ajbc/capsule.

B Additional Results

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

Table 6 lists top documents for an event described in Section 3. Table 7 shows a selection of general topics and Table 8 shows a selection of entity topics.

Model Sensitivity. We assessed the sensitivity of our model to three different decay functions f: exponential, linear, and step functions. We simulated data for each function and then fit Capsule using every permutation of f and multiple settings for event decay duration. We considered a step function,

$$f(i_d, t) = \begin{cases} 1, & \text{if } t \le i_d < t + \tau \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

as well as linear decay,

$$f(i_d, t) = \begin{cases} 1 - \frac{i_d - t}{\tau}, & \text{if } t \le i_d < t + \tau \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

and an exponential decay function:

$$f(i_d, t) = \begin{cases} 0, & \text{if } t \le i_d < t + \tau \\ \exp\left\{\frac{-(i_d - t)}{\tau/5}\right\}, & \text{otherwise.}^{2} \end{cases} \tag{20}$$

Algorithm 1: Variational Inference for Capsule

Input: word counts w
Output: approximate posterior of latent parameters in terms of variational parameters $\boldsymbol{\lambda}$
Initialize $\mathbb{E}[\beta]$ to slightly random around uniform
Initialize $\mathbb{E}$[all other parameters] to uniform
for *iteration* $m = 1 : M$ **do**
  set all $\boldsymbol{\lambda}$ to respective priors, excluding $\lambda^{\theta,rate}$, $\lambda^{\xi,rate}$, 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 $\lambda_v^{\gamma} \mathrel{+}= \mathbb{E}[z_{d,v}^{\mathcal{T}}]$ [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 $\lambda_d^{\epsilon,rate} = \mathbb{E}[\psi] + f \sum_v \mathbb{E}[\gamma]$ [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 $\mathbb{E}[\gamma_t] = \lambda^{\gamma_{t,v}} / \sum_v \lambda^{\gamma_t} \forall t$
end
return λ

$f * \epsilon$	**Date**	**Entity**	**Subject**
6.86	1976-07-07	Cairo	Possible SC meeting on Israeli rescue operation
6.18	1976-07-10	Kuwait	Media reaction to Bicentennial summary
6.15	1976-07-06	Damascus	Syria condemns Israeli operation to free Air France ...
5.91	1976-07-08	Tel Aviv	Passengers comment on Air France hijacking
5.89	1976-07-06	Stockholm	Possible SC meeting on Israeli rescue operation
5.38	1976-07-09	Nicosia	Bicentennial activities in Cyprus
5.09	1976-07-11	State	Security Council debate on Entebbe events CONFID...
4.77	1976-07-09	State	Travel of Peter M. Storm, House Budget Committee
4.76	1976-07-06	Jidda	Weekly Saudi Editorial Summary (June 30-July 6)
4.68	1976-07-08	Lusaka	SWAPO President seeks assessment of Kissinger-Vor...
4.56	1976-07-07	Stockholm	Ugandan role in Air France hijacking
4.45	1976-07-06	Karachi	Transitional quarter funding for RSS travel
4.43	1976-07-06	Athens	Bicentennial anniversary in Greece
4.37	1976-07-08	Damascus	Beirut travel
4.34	1976-07-10	State	Status of Mrs. Bloch
4.17	1976-07-07	Hong Kong	Hong Kong Communist press denounces Israeli resc...
4.12	1976-07-08	Dar es Salaam	President Nyerere's fourth of July messages
4.09	1976-07-10	Moscow	Pravda and Krasnaya Zvezda on Entebbe rescue oper...

Table 6: Top documents for the week after the US bicentennial celebration and Operation Entebbe. Capsule identifies documents relevant to both these real-world events.

We used duration $\tau = 3$ and simulated ten data sets for each of the three functions f. In fitting the models, we also considered all three functions f and varied the decay duration τ from 1 to 5. Figure 6 shows the results of these experiments, using both event detection and document recovery metrics discussed previously.

As expected, the model performs best when the model decay function matches the function used to generate the data. For both event detection and document recovery, the exponential decay was least sensitive to the setting of duration τ used in fitting the data; it was also the least sensitive to the function used in simulating the data. In exploring results on the real-world cable data, we found that the exponential decay provided the most interpretable results.

References

Zoubin Ghahramani and Matthew J Beal. 2001. Propagation algorithms for variational bayesian learning. *Advances in Neural Information Processing Systems (NIPS)*, 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, November.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, January.

[2] Unlike the linear and step functions, the exponential function could be evaluated for any time interval t after a document's appearance at i_d; the function is truncated for computational reasons. The mean lifetime of this exponential decay is the duration τ is divided by 5—this ensures that 99.3% of the area under the curve is reached before the function is truncated at duration τ.

top terms
church, vatican, catholic, bishop, pope, ford, cardinal, ban, religious, archbishop
program, university, grant, education, school, post, institute, research, center, american
security, council, terrorist, threat, sc, sabotage, protective, herein, unsc, honour
visit, hotel, schedule, arrival, arrive, depart, please, meet, day, room
labor, union, strike, ilo, employment, federation, afl cio, trade, worker, confederation
bank, credit, loan, investment, finance, payment, financial, eximbank, opic, central
law, case, court, legal, investigation, arrest, justice, sentence, trial, attorney
party, government, election, opposition, national, leader, campaign, vote, support, anti
tax, company, pay, lease, compensation, exemption, repatriation, income, taxation, fee
oil, petroleum, opec, crude, gulf, price, exploration, refinery, energy, company
israel, arab, israeli, middle, egypt, peace, plo, cairo, egyptian, lebanon
radio, television, broadcast, allotment, appropriation, obligation, zero, warc, transmitter, network
india, indian, pakistan, delhi, goi, ocean, bangladesh, transit, pakistani, afghan
turkish, turkey, cyprus, greek, greece, athens, ankara, morocco, cypriot, algeria
aid, relief, emergency, usaid, disaster, donor, wfp, sahel, ifad, unicef
aircraft, team, flight, clearance, transport, civair, aviation, traffic, charter, cargo
soviet, moscow, press, ussr, soviet union, american, one, war, communist, article
sea, zone, marine, maritime, fish, coastal, continental, territorial, mile, fishery

Table 7: Top vocabulary terms for a selection of general topics, one per row, according to topic distributions β_k. Capsule identifies general diplomatic themes that can be relevant to any entity.

entity	top terms
Ankara	turkish, turkey, ankara, government, cyprus, greek, party, one, time
Athens	greek, athens, greece, gog, government, cyprus, turkish, press, minister
Auckland	new zealand, company, box, trade, contact, opportunity, united states
Baghdad	iraqi, iraq, goi, arab, state, regime, ministry, government, party
Berlin	berlin, frg, german, senat, time, bonn, trade, one, agreement
Bern	swiss, bern, federal, bank, snb, gold, end, interest, national
Brussels	belgian, belgium, brussels, government, firestone, european, ministry
Budapest	hungarian, hungary, trade, mudd, one, time, puja, well, policy
Buenos Aires	argentine, argentina, goa, us, hill, government, one, press, police
Cairo	egyptian, cairo, egypt, arab, israeli, israel, peace, agreement, president
Canberra	australian, australia, goa, government, minister, whitlam, end, dfa, time
Dakar	senegalese, president, african, summary, conference, end, support, one
Dar es Salaam	tangov, salaam, tanzanian, spain, president, government, african, one
Guayaquil	ecuador, ecuadorean, port, congen, one, tuna, local, time, boat
Islamabad	pakistan, gop, government, one, party, minister, general, opposition, ppp
Paris	paris, france, rush, french, one, government, amconsul, quai, european
Jerusalem	jerusalem, bank, israeli, us, israel, plo, one, arab, unifil
Jidda	saudi, jidda, saudi arabia, prince, us, fahd, one, time, government
Johannesburg	black, africa, african, trade, union, police, labor, one, committee
Kabul	afghan, government, goa, minister, one, pakistan, regime, time, ministry
Lima	peru, gop, lima, peruvian, dean, minister, general, marcona, government
Lisbon	portugal, portuguese, gop, lisbon, government, party, summary, minister
London	london, british, government, fco, labor, agreement, one, washdc, summary
Madrid	spanish, spain, madrid, one, govt, general, committee, government, time
Nairobi	kenya, nairobi, marshall, embassy, kenyan, unep, le, ref, state
Oslo	norwegian, norway, soviet, government, minister, ministry, policy
Ottawa	canadian, canada, goc, ottawa, us, extaff, government, minister, federal
Peking	chinese, peking, uslo, china, people, teng, one, trade, delegation, hong
Phnom penh	penh, phnom, khmer, rice, fank, enemy, cambodia, government, dean
Prague	czechoslovak, goc, czech, trade, embassy, one, mfa, time, cssr
Quito	ecuador, ecuadorean, gulf, government, minister, bloomfield, general, one
Sao Paulo	paulo, brazil, state, brazilian, president, government, congen, one, do
Seoul	korea, korean, rok, rokg, seoul, park, government, president, time
Singapore	singapore, asean, minister, government, one, prime, comment, vietnam
Sofia	bulgarian, trade, one, agreement, american, visit, committee, party
Sydney	australia, australian, one, general, american, state, government, post
Tokyo	japan, japanese, tokyo, fonoff, summary, miti, end, diet, time
Taipei	taiwan, groc, china, chinese, government, american, one, local, republic
The Hague	dutch, netherlands, hague, government, minister, party, stoel, mfa, one
USUN New York	committee, usun, priority, report, draft, resolution, sc, comite, rep, new york
Vancouver	canada, government, canadian, british, columbia, pipeline, federal, editorial
Zagreb	yugoslav, yugoslavia, croatian, fair, belgrade, american, one, ina, summary
Zurich	swiss, congen, consulate, general, american, bern, dollar, shipment

Table 8: Top vocabulary terms for a selection of entities according to entity-exclusive topics η_n. Capsule identifies entity-specific themes and interests.

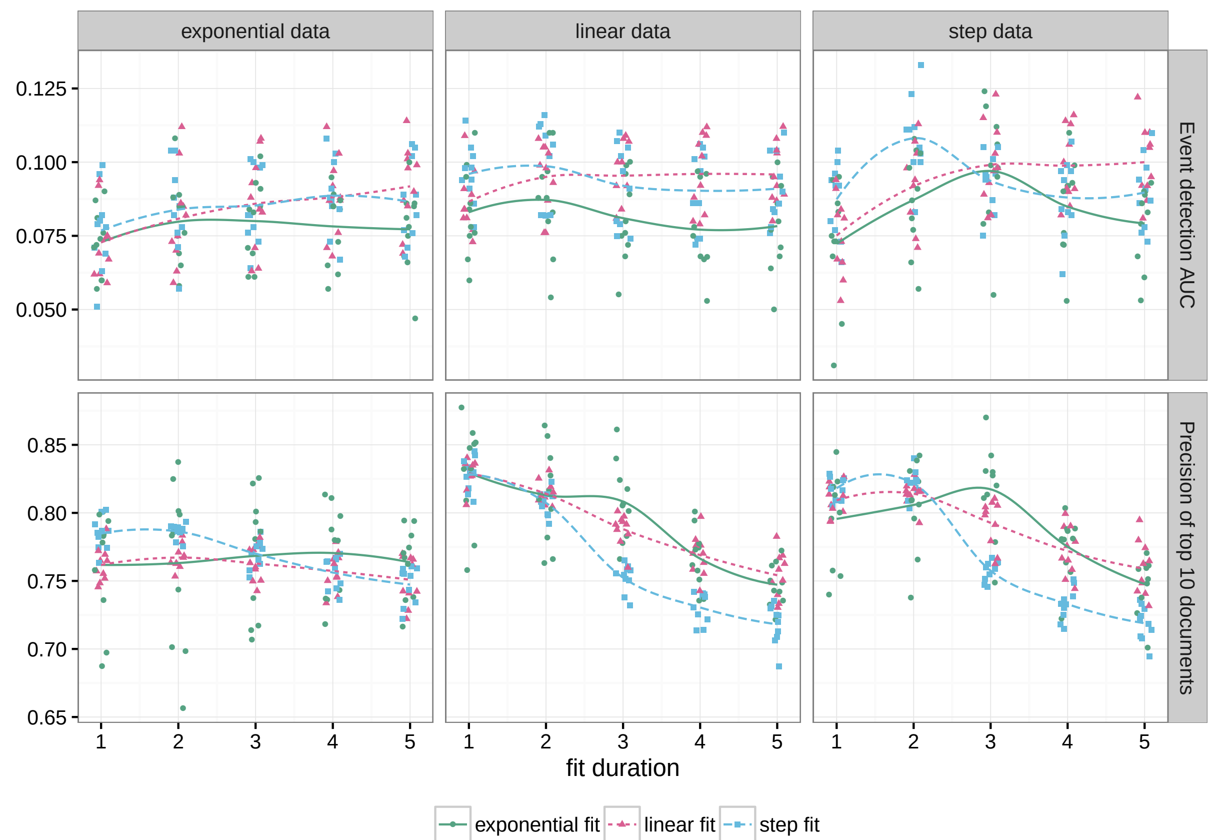

Figure 6: Assessment of model parameter sensitivity on simulated data–Capsule performs best when the model decay function matches the function used to generate the data. The exponential decay is least sensitive to the setting of duration τ and the true function f.