Topic Modeling on Health Journals with Regularized Variational Inference

Robert Giaquinto and Arindam Banerjee, University of Minnesota
{giaquinto.ra@gmail.com, banerjee42@gmail.com}

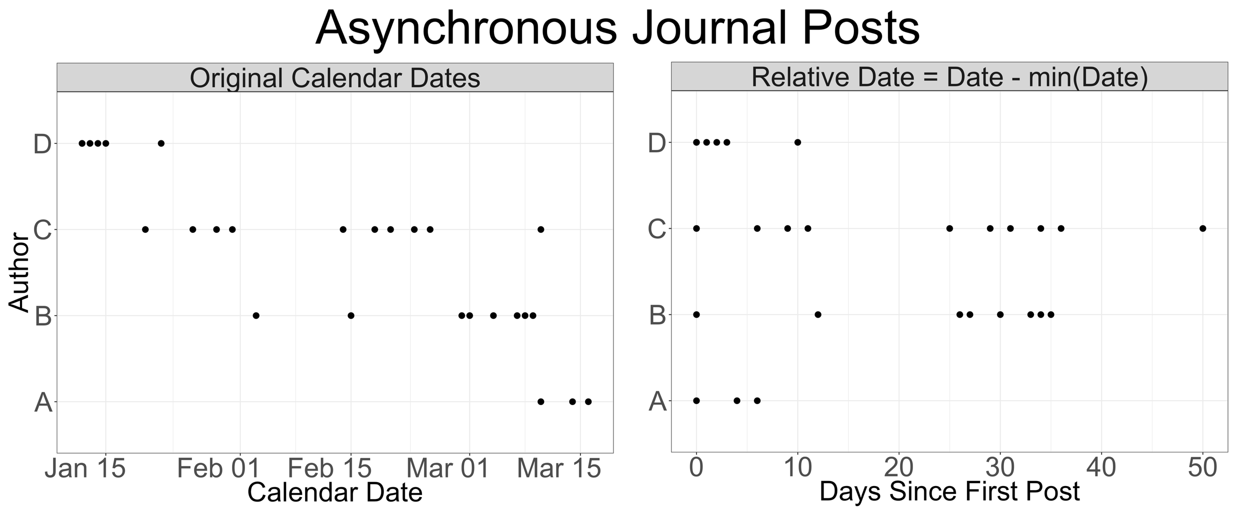
Introduction

- **Objective**: Design a new topic model to meet challenges presented by the CaringBridge (CB) dataset.
  - CB dataset includes patients and caregivers writing during a health crisis.
- **Challenges:** Asynchronous nature of CaringBridge journals.
- **Method:** Develop the **Dynamic Author-Persona (DAP)** topic model for corpora with multiple authors writing over time.
  - Represent authors by a persona – personas capture propensity to write about certain topics over time.
  - Introduce **regularized variational inference (RVI)** algorithm to encourage personas to be distinct.
- **Results:**
  - Better likelihood's compared to competing models.
  - Compelling qualitative results describing common health journeys experienced by CB authors.

Problem

- Want to capture health journeys – clusters of authors with common topic trajectories.
- **Existing methods not adequate**. State-of-the-art topic models [1, 2, 3, 4, 5]:
  - Identify topics, track changes over time to topic word distributions, or associate authors with certain topics.
  - What about common narratives and the authors sharing them?
- Topic model must handle **asynchronous** nature exhibited by CB data.
  - Authors start and stop journaling at different times (in calendar dates and how far along they are in their health journey).
  - Authors post at irregular frequencies (e.g. posting after major event, less often as time progresses).
  - Patients with chronic condition versus brief ailment.

Asynchronous Journal Posts



Data and Evaluation

- **CaringBridge Dataset:** Journals written by patients and caregivers during a health crisis.
- Full dataset includes 13.1 million journals written by approximately 500k authors between 2006 and 2016.
- **Preprocessing:** 1st year of posting, only posts with 10 or more words, authors posting >2 per month.
- **Evaluation set:** 2,000 authors were randomly selected.
  - Total of 114,532 journals.
  - Authors journal an average of 57 times in 1st year (~5 days between journal posts).
- **Evaluation Procedure:** Journals split into training (90%, N=103,018) and test sets (10%, N=11,728).
  - Model variance estimated by 10-fold cross-validation.

Acknowledgements

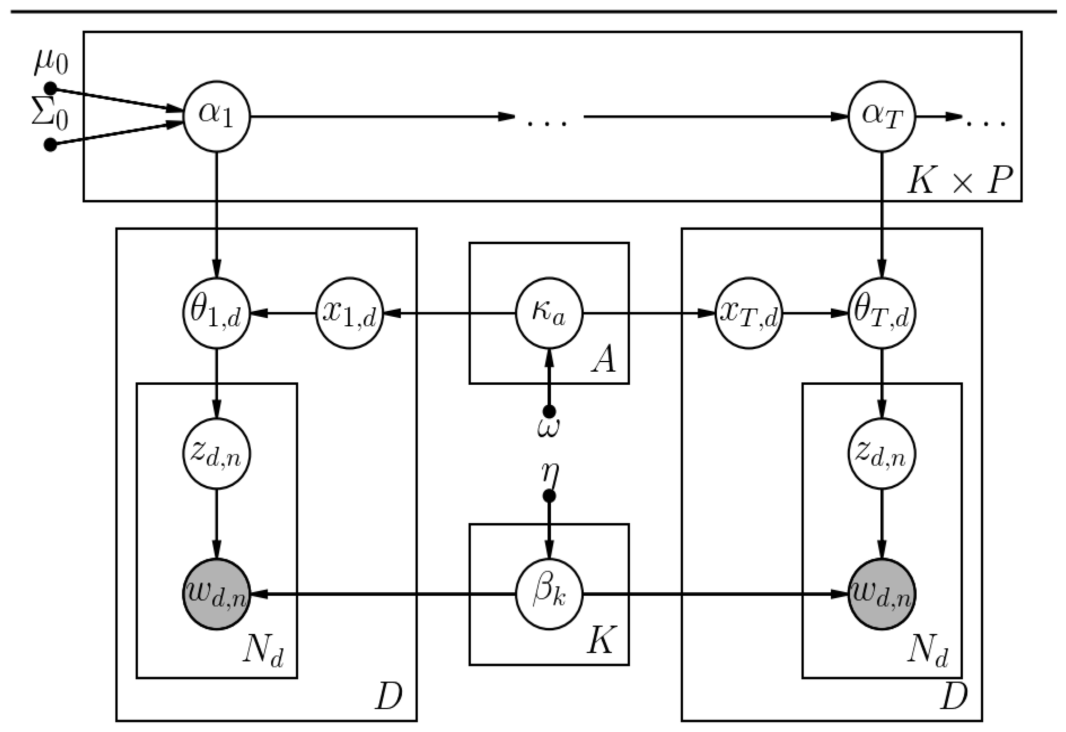
We thank University of Minnesota Supercomputing Institute, and CaringBridge for their support and collaboration. Research supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS-1314560.

References

[1] Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(4-5):993–1022.
[2] Lafferty, J. D., and Blei, D. M. 2006. Correlated Topic Models. *Advances in Neural Information Processing Systems 18* 147–154.
[3] Blei, D. M., and Lafferty, J. D. 2006. Dynamic Topic Models. *International Conference on Machine Learning* 113– 120.
[4] Wang, C.; Blei, D.; and Heckerman, D. 2008. Continuous Time Dynamic Topic Models. *Proc of UAI* 579–586.
[5] Mimno, D., and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* 500–509.
[6] Wainwright, M. J., and Jordan, M. I. 2007. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* 1(1):1–305.

Dynamic Author-Persona Topic Model

Parameter	Variational	Description
$\mathbf{w}_{t,d}$		Words in document d_t
$\mathbf{z}_n$	$\boldsymbol{\phi}_n$	Assigns word n to a topic
$\boldsymbol{\theta}_{t,d}$	$\boldsymbol{\gamma}_{t,d}$	Distribution over topics for d_t
$\mathbf{v}_{t,d}$	$\hat{\mathbf{v}}_{t,d}$	Covariance between topics for d_t
μ_0		Prior for mean of $\boldsymbol{\alpha}_0$
Σ_0		Prior for covariance of $\boldsymbol{\alpha}_0$
$\boldsymbol{\alpha}_t$	$\hat{\boldsymbol{\alpha}}_t$	$\forall p$ distribution over topics
$\boldsymbol{\Sigma}_t$	$\hat{\boldsymbol{\Sigma}}_t$	Covariance in topic distributions
ω		Prior for κ_a
$\boldsymbol{\kappa}_a$	$\boldsymbol{\delta}_a$	Author's distribution over personas
$\mathbf{x}_{d,t}$	$\boldsymbol{\tau}_{t,d}$	Assigns author of d_t to a persona
η		Prior parameter for β_k
$\boldsymbol{\beta}_k$	$\boldsymbol{\lambda}_k$	$\forall k$ distribution over words



Regularized Variational Inference

- **Idea:** Nudge $\hat{\alpha}_{t,p}$ to find different topic distributions for each persona.
- **Approach:** Optimize a penalized surrogate likelihood, i.e. ELBO plus regularization term [6].
- **Regularization term:** inner product between personas (excluding a persona with itself):

$$\rho r(q) = \sum_{p=1}^{P} \sum_{1\leq q\leq P, q\neq p} \frac{D_t}{2}\rho\hat{\alpha}_{t,p}^{\top}\Sigma_t^{-1}\hat{\alpha}_{t,q}\ ,$$

- $\hat{\alpha}_{t,p}$ **Update:** Take gradient w.r.t. $\hat{\alpha}_{t,p}$ of regularization term and $\hat{\alpha}_{t,p}$ terms in ELBO. Update to $\hat{\alpha}_{t,p}$ is:

$$(1+\sum_{d=1}^{D_t}\tau_{d,p}^2)\hat{\alpha}_{t,p} + \rho D_t \sum_{q\neq p}\hat{\alpha}_{t,q} = \hat{\alpha}_{t-1,p} + \sum_{d=1}^{D_t}(\gamma_d - 1)\tau_{d,p}$$

- **Solving for** $\hat{\alpha}_{t,p}$: RHS known, similarly τ on LHS computed during E-step. Therefore, can solve as linear system (Ax=b) where $\hat{\alpha}_{t,p}$ is unknown.

Variational E-Step

- Estimate variational parameters for each document.
- Update to ϕ mimics CTM [2].
- Must estimate τ with exponentiated gradient descent.

$$\frac{\partial\mathcal{L}}{\partial\tau_{t,d,p}} = \Psi(\delta_{a,p}) - \Psi(\sum_{i=1}^{P}\delta_{a,i}) - \log\tau_{a,p} - 1 + \lambda +$$
$$\hat{\alpha}_{t,p}\Sigma_t^{-1}(\gamma_{t,d} - \hat{\alpha}_{t,p}\tau_{t,d,p}) - \frac{1}{2}\mathrm{Tr}(\Sigma_t^{-1}\,\mathrm{diag}(\hat{\alpha}_{t,p}^2 + \hat{\Sigma}_t))$$

- Must estimate each document's topic distribution γ with conjugate gradient descent.

$$\frac{\partial\mathcal{L}}{\partial\gamma_{t,d,k}} = -\Sigma_t^{-1}(\gamma_{t,d,k} - \hat{\alpha}_{t,1:P,k}\tau_{t,d,k}) +$$
$$\sum_{n=1}^{N_{d_t}}\phi_{n,k} - \frac{N_{d_t}}{\zeta}\exp(\gamma_{t,d,k} + \hat{v}_{t,k}^2/2))$$

- Estimate noisy variational observation $\hat{\alpha}_{t,p}$ based on update equations for RVI.

Variational M-Step

- Update global parameters $\boldsymbol{\beta}$, and $\boldsymbol{\kappa}$ in standard fashion, similar to LDA. Simple, closed form.
- Smooth $\hat{\alpha}_{t,p}$ with variational Kalman Filter [4] to estimate $\boldsymbol{\alpha}$

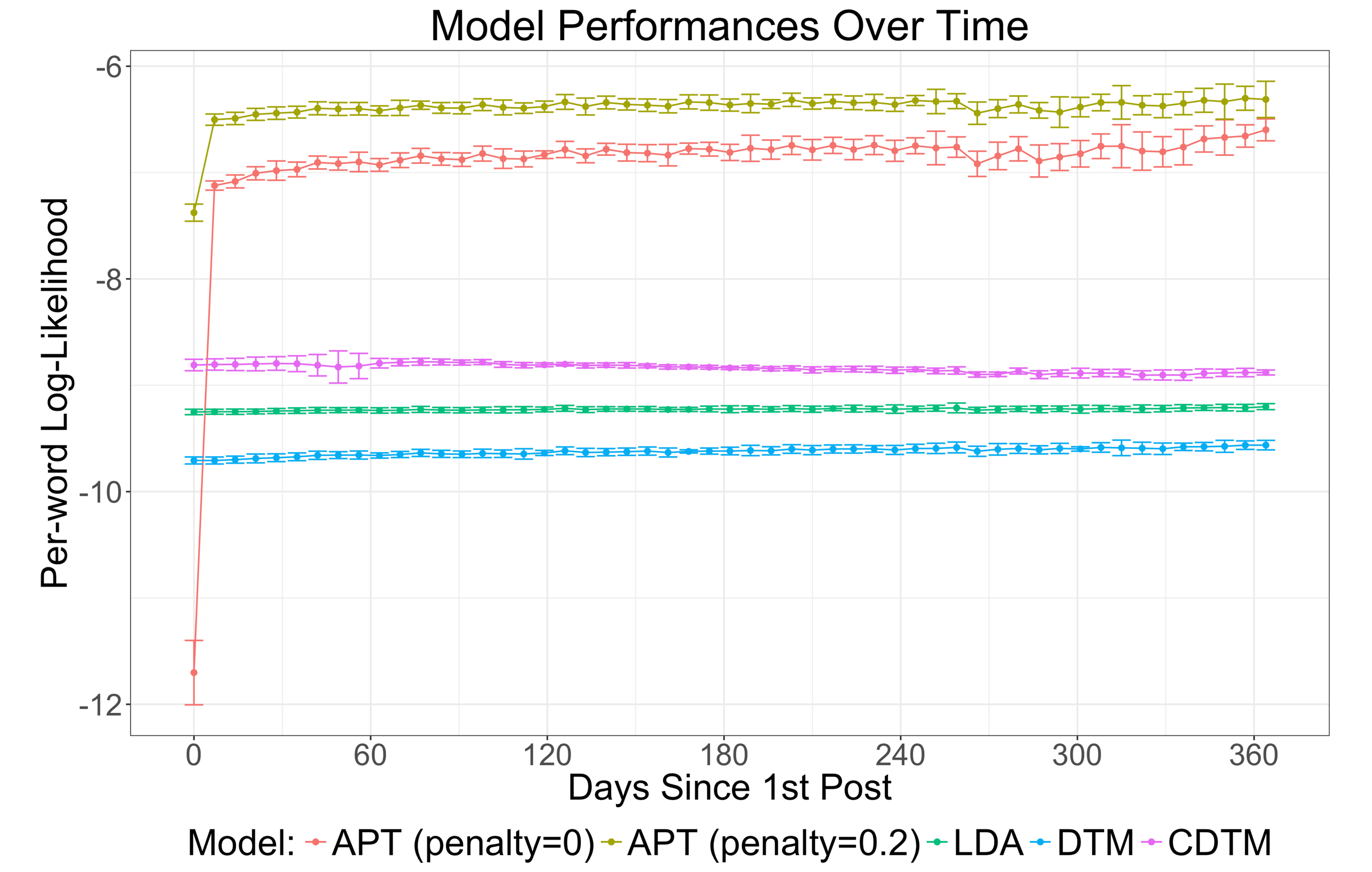
Results

- **Quantitative:** Compare per-word log-likelihoods of documents in test set for DAP, LDA, DTM, and CDTM.

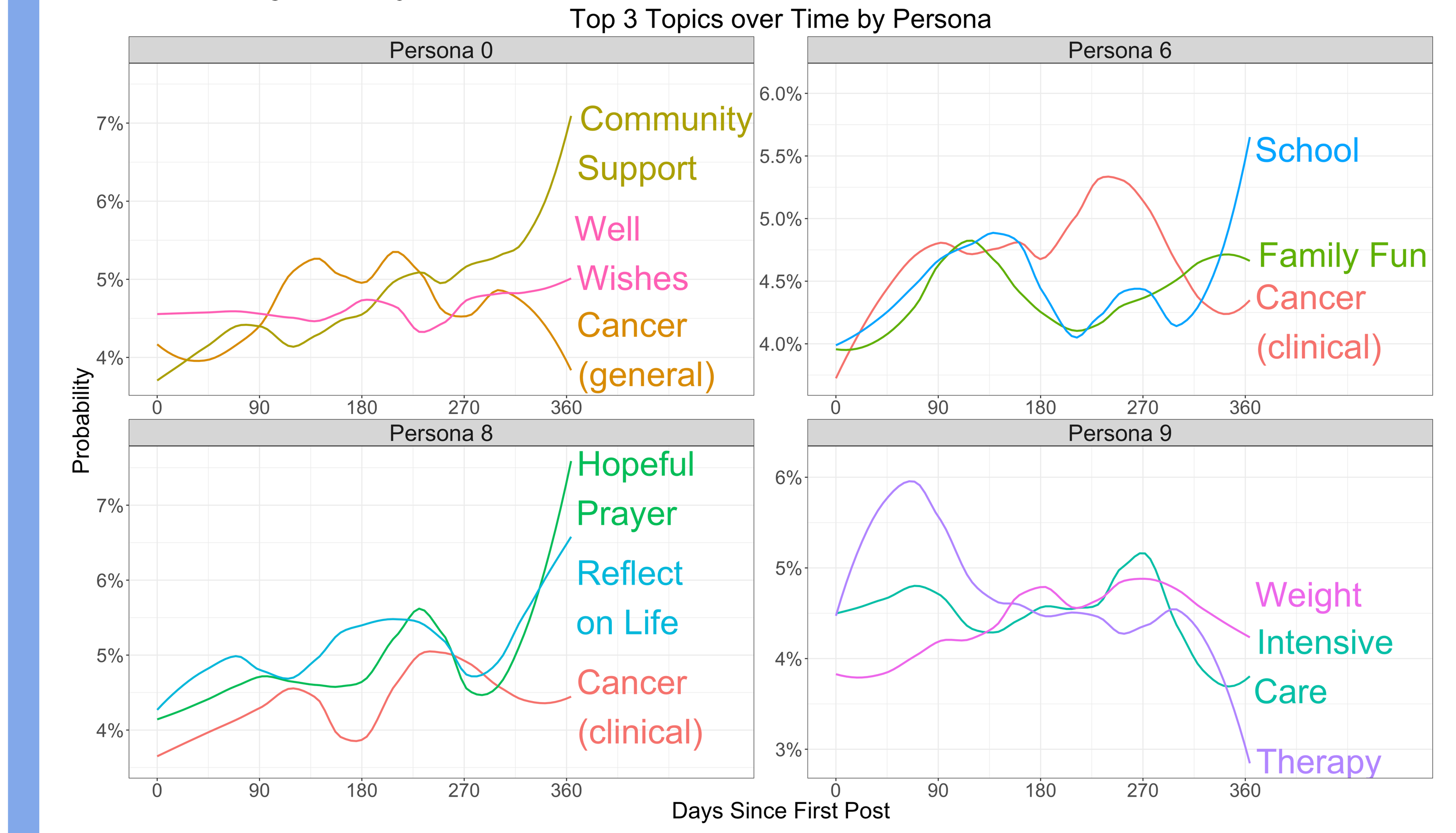
Model	Per-word Log-Likelihood	Std. Dev.
DAP (ρ=0.0)	-7.22	0.04
DAP (ρ=0.2)	**-6.47**	**0.04**
LDA	-9.23	0.02
DTM	-9.65	0.03
CDTM	-8.82	0.03

Results

- DAP model performs better than competing models over time steps. Error bars show one st-dev. in document-level PWLL.



- **Qualitative:** Are personas distinct, and do they capture coherent health journeys?



- DAP finds compelling, unique personas corresponding to common health journeys experienced by CaringBridge users.
- Personas 0 engages with community, and less clinical when writing about cancer.
- Personas 6 and 8 write about cancer using clinical terminology.
  - Persona 6's non-health updates on school, family, celebrations.
  - Persona 8's non-health updates are deep, reflective, prayerful.
- Persona 9 begins with therapy for physical ailment, followed by intensive care and attention to weight.

Community Support	Physical Therapy	Reflect on Life	Hopeful Prayer	Family Fun	Infection	Weather	School
family	therapy	life	god	christmas	blood	nice	school
friend	rehab	know	pray	play	infection	weather	shot
church	therapist	child	prayer	birthday	fluid	walk	go
thank	physical	never	lord	game	fever	lunch	appt
card	pt	love	bless	fun	antibiotic	cold	class
love	chair	year	please	kid	pressure	snow	tomorrow
service	speech	live	heal	party	kidney	outside	grandma
friends	progress	people	trust	year	iv	breakfast	teacher
support	move	cancer	peace	enjoy	lung	rain	home
gift	arm	moment	continue	dinner	clot	go	aunt

Cancer (clinical)	Cancer (general)	Intensive Care	Well Wishes	Hair Loss	Surgery	Bedtime	Weight
chemo	cancer	tube	dad	hair	surgery	sleep	weight
blood	treatment	breathe	mom	leg	surgeon	night	mommy
count	radiation	oxygen	everyone	wear	heart	bed	gain
bone	scan	lung	message	head	dr	wake	feed
marrow	chemo	feed	guestbook	look	office	nurse	daddy
platelet	tumor	x_ray	please	cut	op	say	bottle
round	oncologist	chest	prayer	knee	procedure	asleep	pound
clinic	dr	nurse	read	hat	cardiologist	_time_	feeding
transfusion	ct	vent	visit	wig	valve	room	oz
url	result	stomach	update	shave	ha	tell	milk

Table 2: Top 10 words associated with the most prevalent topics found by the DAP model ($\rho = 0.2$). Topic labels are selected manually in order to aid reference with Figure 3. The words _time_ and _URL_ refer to the result of text pre-processing steps for capturing common patterns like the time of day and website URLs, respectively.

Conclusions

- DAP is uniquely suited to model text data with a temporal structure and written by multiple authors.
- DAP discovers latent personas – a novel component that identifies authors with similar topics trajectories.
- RVI algorithm further improves the DAP model's performance.
- We introduce the CaringBridge dataset: a massive collection of journals written by patients and caregivers, many of who face serious, life-threatening illnesses.
- From the CB dataset DAP extracts compelling descriptions of health journeys.