Evaluation Methods for Topic Models

Hanna M. Wallach

University of Massachusetts Amherst
wallach@cs.umass.edu

April 13, 2009

Joint work with Iain Murray, Ruslan Salakhutdinov and David Mimno

Statistical Topic Models

- Useful for analyzing large, unstructured text collections

bounds	units	policy	data	neurons
bound	hidden	action	space	neuron
loss	network	reinforcement	clustering	spike
functions	layer	learning	points	synaptic
error	unit	actions	distance	firing

- Topic-based search interfaces (`http://rexa.info`)
- Analysis of scientific trends (Blei & Lafferty, '07; Hall et al., '08)
- Information retrieval (Wei & Croft '06)

Latent Dirichlet Allocation (Blei et al., '03)

- LDA generates a new document $\mathbf{w}$ by drawing:

$\boldsymbol{\theta} \sim \mathrm{Dir}\left(\boldsymbol{\theta}; \alpha \mathbf{m}\right)$	a document-specific topic dist.,
$\mathbf{z} \sim P(\mathbf{z} \mid \boldsymbol{\theta}) = \prod_n \theta_{z_n}$	a topic assignment for each token,
$\mathbf{w} \sim P(\mathbf{w} \mid \mathbf{z}, \Phi) = \prod_n \phi_{w_n \mid z_n}$	and finally the observed tokens.

- The "topic" parameters Φ, and $\alpha \mathbf{m}$, are shared by all documents
- For real-world data, only the tokens $\mathbf{w}$ are observed

Evaluating Topic Model Performance

- Unsupervised nature of topic models makes evaluation hard
- There may be extrinsic tasks for some applications...
- ... but we also want to estimate cross-task generalization
- Compute probability of held-out documents under the model
  - Classic way of evaluating generative models
  - Often used to evaluate topic models
- This talk: demonstrate that standard methods for evaluating topic models are inaccurate and propose two alternative methods

Evaluating LDA

- Given training documents $\mathcal{W}'$ and held-out documents $\mathcal{W}$:

$$P(\mathcal{W} \mid \mathcal{W}') = \int \mathrm{d}\Phi \, \mathrm{d}\alpha \, \mathrm{d}\mathbf{m} \, P(\mathcal{W} \mid \Phi, \alpha\mathbf{m}) \, P(\Phi, \alpha\mathbf{m} \mid \mathcal{W}')$$

- Approximate this integral by evaluating at a point estimate
- Variational or MCMC can be used to marginalize out topic assignments for training documents to infer Φ and $\alpha\mathbf{m}$
- The probability of interest is therefore:

$$P(\mathcal{W} \mid \Phi, \alpha\mathbf{m}) = \prod_d P(\mathbf{w}^{(d)} \mid \Phi, \alpha\mathbf{m})$$

Computing $P(\mathbf{w} \mid \Phi, \alpha\mathbf{m})$

- $P(\mathbf{w} \mid \Phi, \alpha\mathbf{m})$ is the normalizing constant that relates the posterior distribution over $\mathbf{z}$ to the joint distribution over $\mathbf{w}$ and $\mathbf{z}$:

$$P(\mathbf{z} \mid \mathbf{w}, \Phi, \alpha\mathbf{m}) = \frac{P(\mathbf{w}, \mathbf{z} \mid \Phi, \alpha\mathbf{m})}{P(\mathbf{w} \mid \Phi, \alpha\mathbf{m})}$$

- Computing it involves marginalizing over latent variables:

$$P(\mathbf{w} \mid \Phi, \alpha\mathbf{m}) = \sum_{\mathbf{z}} \int \mathrm{d}\boldsymbol{\theta}\, P(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta} \mid \Phi, \alpha\mathbf{m})$$

Methods for Computing Normalizing Constants

- Simple importance sampling methods:
  - e.g., MALLET's "empirical likelihood", "iterated pseudo-counts"
- The "harmonic mean" method (Newton & Raftery, '94):
  - Known to overestimate, yet used in topic modeling papers
- Annealed importance sampling (Neal, '01):
  - Accurate, but prohibitively slow for large data sets
- A Chib-style method (Murray & Salakhutdinov, '09)
- A "left-to-right" method (Wallach, '08)

Chib-Style Estimates

- For *any* "special" set of latent topic assignments $\mathbf{z}^\star$:

$$P(\mathbf{w} \,|\, \Phi, \alpha\mathbf{m}) = \frac{P(\mathbf{w} \,|\, \mathbf{z}^\star, \Phi)\, P(\mathbf{z}^\star \,|\, \alpha\mathbf{m})}{P(\mathbf{z}^\star \,|\, \mathbf{w}, \Phi, \alpha\mathbf{m})}$$

- Chib-style estimation:
  1. Pick some special set of latent topic assignments $\mathbf{z}^\star$
  2. Compute $P(\mathbf{w} \,|\, \mathbf{z}^\star, \Phi)\, P(\mathbf{z}^\star \,|\, \alpha\mathbf{m})$
  3. Estimate $P(\mathbf{z}^\star \,|\, \mathbf{w}, \Phi, \alpha\mathbf{m})$
- Can use a Markov chain to estimate $P(\mathbf{z}^\star \,|\, \mathbf{w}, \Phi, \alpha\mathbf{m})$

Markov Chain Estimation

- Stationary condition for a Markov chain:

$$P(\mathbf{z}^\star \,|\, \mathbf{w}, \Phi, \alpha\mathbf{m}) = \sum_{\mathbf{z}} T(\mathbf{z}^\star \leftarrow \mathbf{z})\, P(\mathbf{z} \,|\, \mathbf{w}, \Phi, \alpha\mathbf{m})$$

- Estimate sum using a sequence of states $\mathcal{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(S)}\}$ generated by a Markov chain that explores $P(\mathbf{z} \,|\, \mathbf{w}, \Phi, \alpha\mathbf{m})$

$$\mathbf{z}^{(1)} \xrightarrow[T]{} \mathbf{z}^{(2)} \xrightarrow[T]{} \mathbf{z}^{(3)} \xrightarrow[T]{} \mathbf{z}^{(4)} \xrightarrow[T]{} \mathbf{z}^{(5)}$$

Overestimate of $P(\mathbf{w} \mid \Phi, \alpha\mathbf{m})$

- $P(\mathbf{z}^\star \mid \mathbf{w}, \Phi, \alpha\mathbf{m})$ is unbiased in expectation:

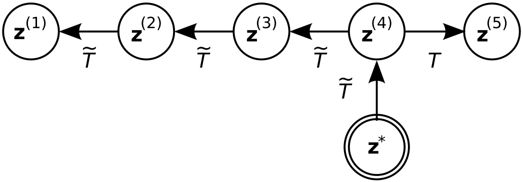
$$P(\mathbf{z}^\star \mid \mathbf{w}, \Phi, \alpha\mathbf{m}) = \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} T(\mathbf{z}^\star \leftarrow \mathbf{z}^{(s)})\right]$$

- But, in expectation, $P(\mathbf{w} \mid \Phi, \alpha\mathbf{m})$ will be *overestimated* (Jensen):

$$\begin{aligned} & P(\mathbf{w} \mid \Phi, \alpha\mathbf{m}) \\ & = \frac{P(\mathbf{z}^\star, \mathbf{w} \mid \Phi, \alpha\mathbf{m})}{\mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} T(\mathbf{z}^\star \leftarrow \mathbf{z}^{(s)})\right]} \leq \mathbb{E}\left[\frac{P(\mathbf{z}^\star, \mathbf{w} \mid \Phi, \alpha\mathbf{m})}{\frac{1}{S}\sum_{s=1}^{S} T(\mathbf{z}^\star \leftarrow \mathbf{z}^{(s)})}\right] \end{aligned}$$

Chib-Style Method (Murray & Salakhutdinov, '09)

- Draw $\mathcal{Z} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(S)}\}$ from a carefully designed distribution



Unbiased: $P(\mathbf{w} \mid \Phi, \alpha\mathbf{m}) \simeq P(\mathbf{w}, \mathbf{z}^\star \mid \Phi, \alpha\mathbf{m}) \Big/ \frac{1}{S} \sum_{s'=1}^{S} T(\mathbf{z}^\star \leftarrow \mathbf{z}^{(s')})$

Left-to-Right Method (Wallach, '08)

- Can decompose $P(\mathbf{w} \mid \Phi, \alpha\mathbf{m})$ as

$$\begin{aligned} P(\mathbf{w} \mid \Phi, \alpha\mathbf{m}) &= \prod_n P(w_n \mid \mathbf{w}_{<n}, \Phi, \alpha\mathbf{m}) \\ &= \prod_n \sum_{\mathbf{z}_{\leq n}} P(w_n, \mathbf{z}_{\leq n} \mid \mathbf{w}_{<n}, \Phi, \alpha\mathbf{m}) \end{aligned}$$

- Approximate each sum over $\mathbf{z}_{\leq n}$ using a MCMC algorithm
- "Left-to-right": appropriate for language modeling applications

Left-to-Right Method (Wallach, '08)

for each position n in $\mathbf{w}$ **do**
  for each particle $r = 1$ to R **do**
    for each position $n' < n$ **do**
      resample $z_{n'}^{(r)} \sim P(z_{n'}^{(r)} \mid w_{n'}, \{\mathbf{z}_{<n}^{(r)}\}_{\setminus n'}, \Phi, \alpha\mathbf{m})$
    end for
    $p_n^{(r)} := \sum_t P(w_n, z_n^{(r)} = t \mid \mathbf{z}_{<n}^{(r)}, \Phi, \alpha\mathbf{m})$
    sample a topic assignment: $z_n^{(r)} \sim P(z_n^{(r)} \mid w_n, \mathbf{z}_{<n}^{(r)}, \Phi, \alpha\mathbf{m})$
  end for
  $p_n := \sum_r p_n^{(r)} \,/\, R$
  $l := l + \log p_n$
end for

Relative Computational Costs

- Gibbs sampling dominates cost for most methods

Method	Parameters	Cost
Iterated pseudo-counts	# itns. I, # samples S	$(I + S)\,N$
Empirical likelihood	# samples S	SN
Harmonic mean	burn-in B, # samples S	$N\,(B + S)$
AIS	# temperatures S	SN
Chib-style	chain length S	$2SN$
Left-to-right	# particles R	$RN\,(N - 1)\,/\,2$

- Costs are in terms of # Gibbs site updates required (or equivalent)

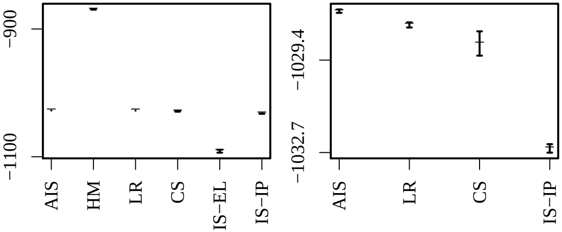
Data Sets

- Two synthetic data sets, three real data sets:

Data set	V	$\bar{N}$	St. Dev.
Synthetic, 3 topics	9242	500	0
Synthetic, 50 topics	9242	200	0
20 Newsgroups	22695	120.4	296.2
PubMed Central abstracts	30262	101.8	49.2
New York Times articles	50412	230.6	250.5

- V is the vocabulary size, $\bar{N}$ is the mean document length, "St. Dev." is the estimated standard deviation in document length

Average Log Prob. Per Held-Out Document (20 Newsgroups)



- **AIS:** Annealed importance sampling. **HM:** Harmonic mean. **LR:** Left-to-right. **CS:** Chib-style. **IS-EL:** Importance sampling (empirical likelihood). **IS-IP:** Importance sampling (iterated pseudocounts)

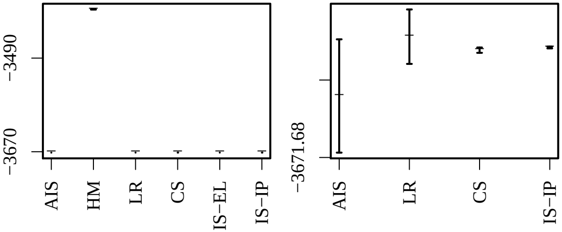
Conclusions

- Empirically determined that the evaluation methods currently used in the topic modeling community are inaccurate:
  - Harmonic mean method often significantly overestimates
  - Simple IS methods tend to underestimate (but not by as much)
- Proposed two, more accurate, alternatives
  - A Chib-style method (Murray & Salakhutdinov, '09)
  - A left-to-right method (Wallach, '08)

Questions?

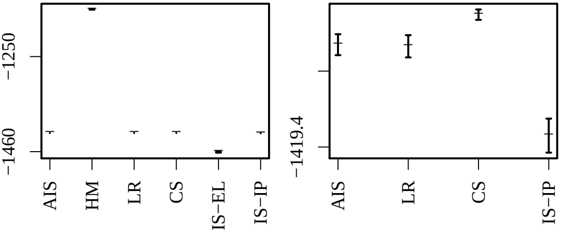
wallach@cs.umass.edu
http://www.cs.umass.edu/~wallach/

Average Log Prob. Per Held-Out Document (Synth., 3 Topics)



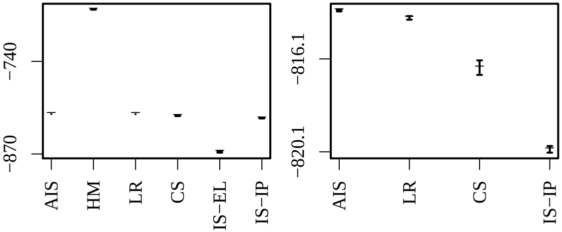
- **AIS:** Annealed importance sampling. **HM:** Harmonic mean. **LR:** Left-to-right. **CS:** Chib-style. **IS-EL:** Importance sampling (empirical likelihood). **IS-IP:** Importance sampling (iterated pseudocounts)

Average Log Prob. Per Held-Out Document (Synth., 50 Topics)



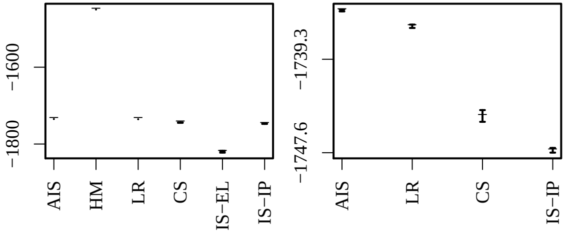
- **AIS:** Annealed importance sampling. **HM:** Harmonic mean. **LR:** Left-to-right. **CS:** Chib-style. **IS-EL:** Importance sampling (empirical likelihood). **IS-IP:** Importance sampling (iterated pseudocounts)

Average Log Prob. Per Held-Out Document (PubMed Central)



- **AIS:** Annealed importance sampling. **HM:** Harmonic mean. **LR:** Left-to-right. **CS:** Chib-style. **IS-EL:** Importance sampling (empirical likelihood). **IS-IP:** Importance sampling (iterated pseudocounts)

Average Log Prob. Per Held-Out Document (New York Times)



- **AIS:** Annealed importance sampling. **HM:** Harmonic mean. **LR:** Left-to-right. **CS:** Chib-style. **IS-EL:** Importance sampling (empirical likelihood). **IS-IP:** Importance sampling (iterated pseudocounts)

Choosing a “Special” State $\mathbf{z}^\star$

- Run regular Gibbs sampling for a few iterations
- Iteratively maximize the following quantity:

$$
\begin{aligned}
&P(z_n = t \,|\, \mathbf{w}, \mathbf{z}_{\setminus n}, \Phi, \alpha\mathbf{m}) \\
&\qquad \propto P(w_n \,|\, z_n = t, \Phi)\, P(z_n = t \,|\, \mathbf{z}_{\setminus n}, \alpha\mathbf{m}) \\
&\qquad \propto \phi_{w_n|t}\, \frac{\{N_t\}_{\setminus n} + \alpha m_t}{N - 1 + \alpha},
\end{aligned}
$$

- $\{N_t\}_{\setminus n}$ is # times topic t occurs in $\mathbf{z}$ excluding position n