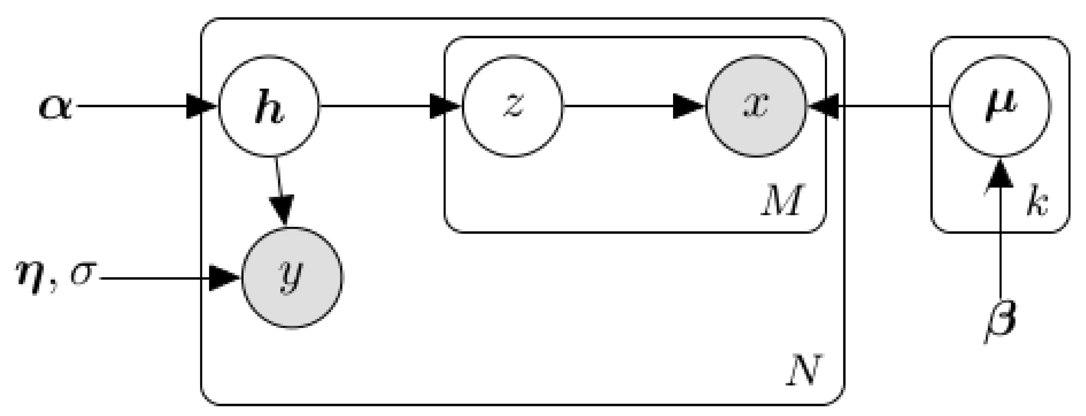
Spectral Methods for Supervised Topic Models

Yining Wang and Jun Zhu

Machine Learning Department, Carnegie Mellon University and TNList, Tsinghua University

Background of sLDA and tensor decomposition

Supervised LDA (Blei and McAuliffe, 2007)



(a) $y_d = \boldsymbol{\eta}_d^\top \boldsymbol{h}_d + \varepsilon_d$

- Data: Document collection $\boldsymbol{x}_{dn}$, response variables $y_d \in \mathbb{R}$.
- $\mathbf{O} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k) \in \mathbb{R}^{V \times k}$: the topic dictionary.
- $\boldsymbol{h}_d \in \Delta^{k-1}$: topic mixing vectors for each document.
- $z_{dn} \in \{1, 2, \cdots k\}$: topic assignments for each word.
- $\boldsymbol{\eta} \in \mathbb{R}^k$: the linear regression model.
- $\boldsymbol{\alpha} \in \mathbb{R}^k$: prior parameter for topic mixing vectors.

Orthogonal tensor decomposition

- For a p-th order tensor $\mathbf{T}$, find orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^r$ and scalars $\{\lambda_i\}_{i=1}^r$ such that $\mathbf{T} = \sum_{i=1}^r \lambda_i \boldsymbol{v}_i^{\otimes p}$.
- Tools: robust tensor power method (Anandkumar et al., 2012), ALS method.

Question: Can we use tensor decomposition based methods to obtain consistent estimates of sLDA parameters?

Key techniques

Observable moments

- $\mathbf{M}_1 = \frac{1}{\alpha_0}\sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i$; $\mathbf{M}_2 = \frac{1}{\alpha_0(\alpha_0+1)}\sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top$;
- $\mathbf{M}_3 = \frac{2}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i^{\otimes 3}$;
- $\mathbf{M}_y = \frac{2}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\sum_{i=1}^k \alpha_i y_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top$.

Simultaneous diagonalization:

- First, find $\mathbf{W}$ that whitens $\mathbf{M}_2$.
- Obtain $\boldsymbol{v}_i = \mathbf{W}\boldsymbol{\mu}_i$ via orthogonal tensor decomposition on whitened tensor $\mathbf{M}_3(\mathbf{W}, \mathbf{W}, \mathbf{W})$.
- Recover $\boldsymbol{\mu}_i$ by multiplying $\mathbf{W}^\dagger$.

Power update for $\boldsymbol{\eta}$: $\eta_i \approx \boldsymbol{v}_i^\top \mathbf{M}_y(\mathbf{W}, \mathbf{W})\boldsymbol{v}_i$.

The learning algorithm and sampling complexity analysis

1. **Input**: Document collection $\boldsymbol{x}_{dn}$, response variables y_d and $\alpha_0 = \|\boldsymbol{\alpha}\|_1$.
2. Compute empirical moments and obtain $\widehat{\mathbf{M}}_2, \widehat{\mathbf{M}}_3$ and $\widehat{\mathbf{M}}_y$.
3. Find $\widehat{\mathbf{W}} \in \mathbb{R}^{n \times k}$ such that $\widehat{\mathbf{M}}_2(\widehat{\mathbf{W}}, \widehat{\mathbf{W}}) = \mathbf{I}_k$.
4. Find robust eigenvalues and eigenvectors $(\widehat{\lambda}_i, \widehat{\boldsymbol{v}}_i)$ of $\widehat{\mathbf{M}}_3(\widehat{\mathbf{W}}, \widehat{\mathbf{W}}, \widehat{\mathbf{W}})$ using the robust tensor power method (Anandkumar et al., 2012).
5. Recover parameters: $\widehat{\alpha}_i \leftarrow \frac{4\alpha_0(\alpha_0+1)}{(\alpha_0+2)^2\widehat{\lambda}_i^2}$, $\widehat{\boldsymbol{\mu}}_i \leftarrow \frac{\alpha_0+2}{2}\widehat{\lambda}_i(\widehat{\mathbf{W}}^+)^\top \widehat{\boldsymbol{v}}_i$, $\widehat{\eta}_i \leftarrow \frac{\alpha_0+2}{2}\widehat{\boldsymbol{v}}_i^\top \widehat{\mathbf{M}}_y(\widehat{\mathbf{W}}, \widehat{\mathbf{W}})\widehat{\boldsymbol{v}}_i$.
6. **Output:** $\widehat{\boldsymbol{\eta}}$, $\widehat{\boldsymbol{\alpha}}$ and $\{\widehat{\boldsymbol{\mu}}_i\}_{i=1}^k$.

Theorem (Sample complexity analysis)

Suppose we have an sLDA model $\mathcal{M} = (\{\boldsymbol{\mu}_i\}_{i=1}^k, \boldsymbol{\alpha}, \boldsymbol{\eta})$ and n documents i.i.d. sampled from $\mathcal{M}$. Let $\mathbf{O} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k)$. Then the estimates $\widehat{\mathcal{M}} = (\{\widehat{\boldsymbol{\mu}}_i\}_{i=1}^k, \widehat{\boldsymbol{\alpha}}_{i=1}^k, \widehat{\boldsymbol{\eta}})$ is ϵ-close to $\mathcal{M}$ with high probability given that $n = \Omega(poly(\epsilon, k, \sigma_k(\mathbf{O})^{-1}, \alpha_{\max}, \alpha_{\min}^{-1}, \|\boldsymbol{\eta}\|))$.

Experiments

Synthetic datasets

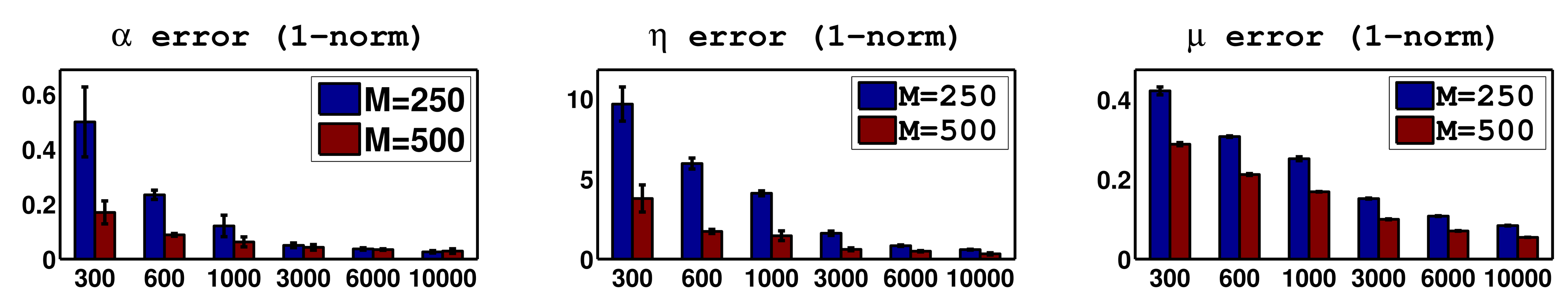


Figure : Parameter estimation error (1-norm) on synthetic datasets.

Amazon movie review dataset (McAuley and Leskovec, 2013)

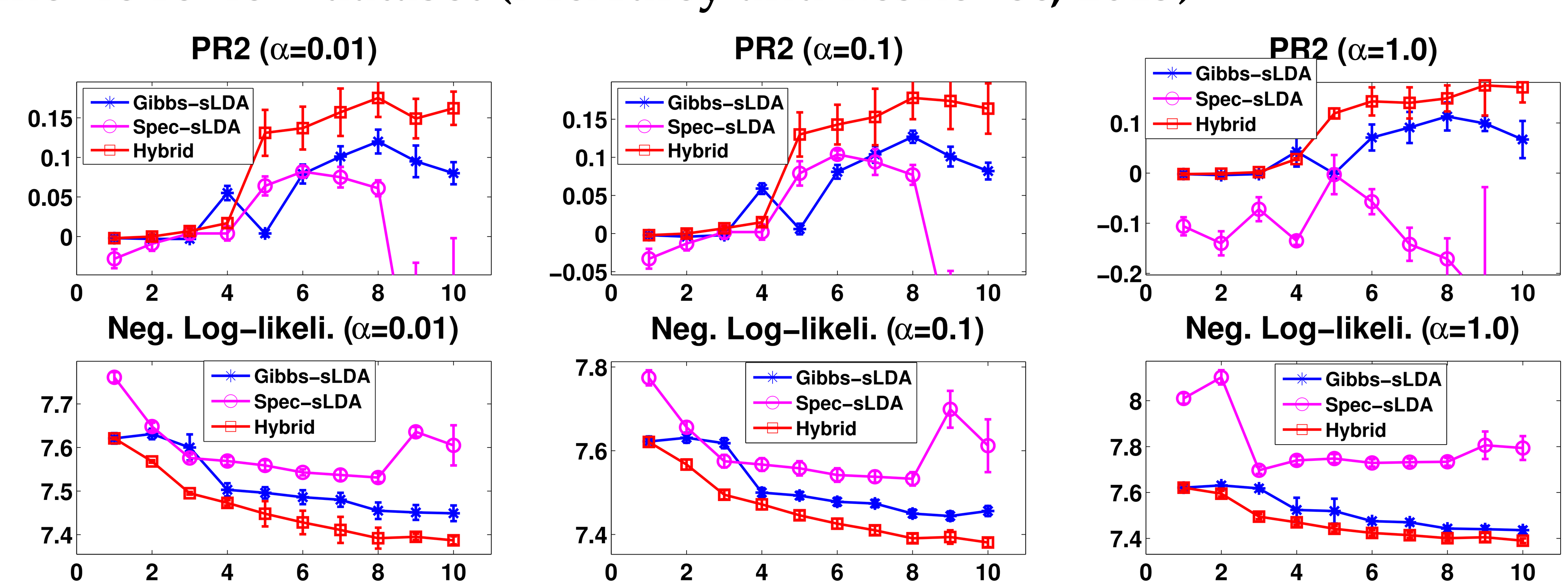


Figure : pR^2 scores (top) and negative per-word log-likelihood (bottom) on Amazon movie review dataset.

Table : Running time for Gibbs sampling and spectral method.

	$k=10$					$k=50$				
$n(\times 10^4)$	1	5	10	50	100	1	5	10	50	100
Gibbs-sLDA	0.6	3.0	6.0	30.5	61.1	2.9	14.3	28.2	145.4	281.8
Spec-sLDA	1.5	1.6	1.7	2.9	4.3	3.1	3.6	4.3	9.5	16.2

yiningwa@cs.cmu.edu dcszj@mail.tsinghua.edu.cn