Scalable Sparse Bayesian Nonparametric and Matrix Tri-factorization Models for Text Segmentation, Topical analysis and Entity Resolution in Dyadic data.

Ranganath B.N.[1] and Shalabh Bhatnagar[1]

[1]Indian Institute of Science, Bangalore, India

April. 28th, 2016

Overview of the work

- To model sparsity for different applications
  - Text segmentation, topical analysis and entity resolution in dyadic data
  - Traditional approaches do not consider sparsity.
- Sparse relationships between the domain entities and the latent components of the data.
- Modeling sparse relationships
  - Extracts relevant information.
  - Scalable solution.
- Bayesian Nonparametric and matrix trifactorization approaches

Block Exchangeable model (BEM):Segmentation

- Segmentation of collections of sequence data
  - Laptop review dataset in to facets
  - News dataset in to stories.
- Occurence of facets or the stories
  - Order independent
  - Persistence according to popularity
  - Different documents share facets or stories
- Exchangeability in the context of segmentation
  - Specifies the type of permutations of assignments to the random variables that do not affect joint probability under a model.
  - Complete or Group Exchangeability
    - All permutations of assignments are equiprobable
    - Segmentation is arbitrary
    - LDA [1] and Hierarchical Dirichlet processes (HDP) [2]: No Persistence aspect
    - Parameters linear in number of documents
  - Markov Exchangeability
    - All permutations with same transition count b/w states are equiprobable.
    - No Order independence though distinction between permutations.
    - HDP-HMM and Sticky HDP-HMM [3]: Inference process expensive
    - Quadratic number of transition parameters

[1] DM Blei et al, Latent dirichlet allocation, JMLR 2003

[2] Teh et al, Hierarchical Dirichlet Processes, JASA 2006

[3] Fox et al, An HDP-HMM for systems with state persistence, ICML 2008

BEM:Solution

- Block Exchangeability (BE) [4]
    - All permutations preserving the same number of self-transitions and number of nonself transitions for each state along with the ending state are equiprobable.
    - Amenable to segmentation, Preserves Order Independence.
    - Sparser number of transition parameters (linear).
    - Scalable Inference
    - BE superclass of CE and subclass of ME.
- Core part of BEM

$$G \sim DP(\delta, H), P_i^* \sim Beta(a, b) \quad for \quad i = 1 : \infty$$
$$\pi_{ij} = (1 - P_i^*)G(\phi_j) + P_i^* \delta(i, j), Z_{ij} \sim (\pi_{(Z_{i(j-1)})})$$

- Inference
    - Intractable due to coupling of P^* and G.
    - Decoupled with introduction of Persistence indicator variables C
        - $C_{ij} = 0/1$, continuity of previous state/ new state
        - Equivalent generative process with collapsed Gibbs sampling
        - Sample Z_{ij} only when $C_{ij} = 1$

[4] Adway et al, A Layered Dirichlet Process for Hierarchical Segmentation of Sequential Grouped Data, ECML 2013

BEM:Experiments

- News Dataset: 150 News transcripts- each having about 18 stories over 5 broad categories-politics,national affairs, international affairs, business and sports.

Model	$fC1$	IT	$Perplexity$	Avg_s
Sticky HDP-HMM	1.0	5.4 sec	3141	**0.3**
BEM1	0.0012	**0.02** sec	3296	0.55
BEM2	0.006	0.05 sec	1988	0.42
BEM3	0.12	0.3 sec	**1103**	0.32

- Laptop reviews: 11000 reviews, each discussing product facets (appearance, weight, connectivity etc)

Model	$fC1$	IT	$Perplexity$	Avg_s
Sticky HDP-HMM	1.0	53 sec	414	**0.27**
BEM1	0.08	**2.1** sec	591	0.57
BEM2	0.19	5.2 sec	**258**	0.3
BEM3	0.22	6.3 sec	299	0.33

Sparse Matrix Tri-factorization (SMTF):Motivation

- Topical analysis of dyadic data
  - Product review data, movie data and bibliographic data.
  - Genres, research and technology areas, product categories are topics in movies, research domain and the product review data.
  - Dyadic association
    - Movie scripts, product literature, research papers and the corresponding words.
    - Consumers, viewers, researchers and the movie scripts, product literature, research papers.
- Sparse relationships between topics and the domain entities
  - Topics and documents
  - Topics and users
- Existing models
  - Probabilistic Author topic model (ATM) [5]
    - Estimates three associations of interest to us
    - Do not address sparsity
  - Collective matrix factorization (CMF) [6]
    - Uses coupled sparse bi-factorization approach.
    - Factorizes a binary author-document association matrix.
    - May not result in accurate estimation of factor matrices
    - Do not estimate the strength of author-document associations.

---

[5] M Rosen-Zvi et al, The author-topic model for authors and documents, UAI 2004

[6] M Sachan et al, Collective matrix factorization for co-clustering, WWW 2013

SMTF:Formulation

$$1, \arg \min_{\Phi,\Theta,\mathbb{A}} \frac{1}{2}\|D - \Phi\Theta\mathbb{A}\|_F^2 + \lambda_1 \sum_{j=1}^{n} \|(\Theta\mathbb{A})_j\|_1 + \lambda_2 \sum_{j=1}^{m} \|\Theta_j\|_1 + \lambda_3 \sum_{j=1}^{t} \|\Phi_j\|_1$$

$$\text{s.t. } supp(\mathbb{A}) \subseteq supp(A), \Phi \geq 0, \Theta \geq 0, \mathbb{A} \geq 0,$$

$$2, \arg \min_{\Phi,\Theta,Q,\mathbb{A}} \frac{1}{2}\|D - \Phi Q\|_F^2 + \frac{1}{2}\|Q - \Theta\mathbb{A}\|_F^2 + \lambda_Q \sum_{j=1}^{n} \|Q_j\|_1 + \lambda_\Theta \sum_{j=1}^{m} \|\Theta_j\|_1$$

$$+\lambda_\Phi \sum_{j=1}^{t} \|\Phi_j\|_1 s.t. supp(\mathbb{A}) \subseteq supp(A), \quad \Phi \geq 0, \Theta \geq 0, Q \geq 0, \mathbb{A} \geq 0,$$

$$2.1, \textit{Solving} \quad \Phi : \quad \{\arg \min_{(\Phi^T)_i \geq 0} \frac{1}{2}\|(D^T)_i - Q^T(\Phi^T)_i\|_F^2 + \lambda_\Phi \|(\Phi^T)_i\|_1\}_{i=1:v}$$

$$2.2, \textit{Solving} \quad \Theta : \quad \{\arg \min_{(\Theta^T)_i \geq 0} \frac{1}{2}\|(Q^T)_i - \mathbb{A}^T(\Theta^T)_i\|_F^2 + \lambda_\Theta \|(\Theta^T)_i\|_1\}_{i=1:t},$$

$$2.3, \textit{Solving} \quad Q : \quad \{\arg \min_{Q_i \geq 0} \frac{1}{2}\|[D; \Theta\mathbb{A}]_i - [\Phi; eye(t,t)]Q_i\|_F^2 + \lambda_Q \|Q_i\|_1\}_{i=1:n}$$

$$2.4, \textit{Solving} \quad \mathbb{A} : \quad \{\arg \min_{\mathbb{A}_j \geq 0} \frac{1}{2}\|Q_j - \Theta\mathbb{A}_j\|_F^2, \mathbb{A}_{ij} = 0 \; \forall i \text{ s.t. } \mathbb{A}_{ij} \notin supp(A)\}_{j=1:n}$$

- In SMTF [7], subproblems are solved in an alternating minimization framework using Projected FISTA.
- Proof for convergence for Projected FISTA.

[7] Ranganath B N, Sparse topical analysis of dyadic data using matrix tri-factorization, Machine Learning 2015

SMTF:Results

- Datasets
  - **DBLP** abstracts dataset (DBLP): We use a subset of 6320 documents involving 3377 authors covering 8 conferences.
  - **Product Review** (REV) dataset: 9651 reviews written by 5675 reviewers in 10 different product categories. This results in 5998 documents, one for each product.
- Sparsity on Q improves TE, DTa-F1, DTa-ARI.
- Sparsity on Q and Θ improves ATa-CCD.

Table: Performance comparison between SMTF, CMF and ATM

	DBLP				REV			
Model	TE	DTa-F1	DTa-ARI	ATa-CCD	TE	DTa-F1	DTa-ARI	ATa-CCD
$SMTF_1$	**131240**	0.33	0.19	**423**	**43129**	0.53	0.46	**454**
$SMTF_2$	150870	0.28	0.13	**423**	48376	0.51	0.44	508
CMF	154580	0.25	0.01	428	48588	0.16	0.01	884
ATM	-	**0.35**	**0.24**	518	-	**0.60**	**0.55**	612

Sparse Entity Resolution Model (SERM):Motivation

- Entity resolution in Dyadic data
  - Identify correct author entity for each of the aliases in all the documents
  - Grouping over the documents
    - Documents share author entities, topics
- Sparse relationships
  - Smaller number of author entities and topics per group.
  - Smaller number of aliases for author entities.
- Existing models
  - LDA for Entity resolution model (LDA-ER) [8]
    - Do not utilize textual information for disambiguation of the identical aliases.
    - Do not address sparsity issue.
  - Grouped Author topic model (GATM) [9]
    - Utilizes textual information.
    - Uses HDP nonparametric prior over author entities and topics for the groups.
    - Do not address sparsity issue.

---

[8] Indrajit Bhattacharya et al, A Latent Dirichlet Model for Unsupervised Entity Resolution, SDM 2006

[9] AM Dai et al, The grouped author-topic model for unsupervised entity resolution, ICANN 2011

SERM:Model

- SERM [10], structurally similar model as that of GATM
- Third level: Stick breaking prior of DP
  - Nonparametric in the number of groups
  - Assignment of groups to the documents.
- Second level: Sparsity promoting Indian Buffet process (IBP) priors over author entities and topics for the groups
  - Nonparametric IBP over author entities for the groups
  - Parametric IBP over topics for the groups leading to scalable solution.
  - Assignment of author entities and topics for the aliases and the words in the group.
- Third level
  - k-NN mechanism for selecting smaller number of potential author aliases for the author entities leading to scalability
  - Noise model as that in LDA-ER, GATM employed for generating the aliases.
  - Generate the word from the topic.

[10] Ranganath B N, Shalabh Bhatnagar, Scalable Focussed Entity Resolution, IJCNN 2016

SERM:Results

- Datasets
  - Citeseer: 1785 author references to the 1009 author entities in 877 documents.
  - Rexa: 2149 author references, among which 747 are labeled and the remaining are unlabeled. The labeled author references point to the 100 author entities in 488 documents.

Table: Best B-CUBED results for SERM and GATM

	Citeseer		Rexa	
Model	time	F1	time	F1
SERM	**3.2**	**86.06**	**1.3**	**77.65**
GATM	21.6	82.21	14	61.49

Questions ?