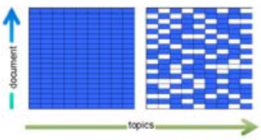
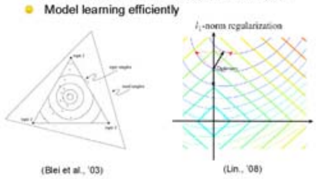
Group Sparse Topical Coding: From Code to Topic

Lu Bai, Jiafeng Guo, Yanyan Lan, Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences
{bailu, guojiafeng}@software.ict.ac.cn, {lanyanyan, cxq}@ict.ac.cn

1. MOTIVATION

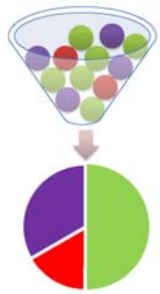


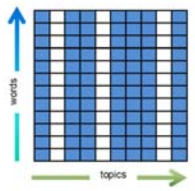
- PTM(Probabilistic Topic Model)
  - Document is modeled as a meaningful low dimensional point in the topic simplex
  - Lack a mechanism to directly control the posterior sparsity of the inferred representations because the normalization constraint
- NPM(Non-Probabilistic Model)
  - Easy to achieve sparsity by using sparse constraint like lasso or other composite regularizer
  - Lose the clear semantic explanation over the latent representations.
- Group Sparse Topical Coding(GSTC)
  - Produce a meaningful representation of document
  - Control the sparsity of representation directly
  - Model learning efficiently



2. BASIC IDEA

- The meaning of document is composed of the meanings of words.
- Restricting the topics of words in the document can impact the meanings of document implicitly.





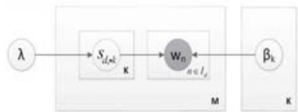
Modeling the word count w_n in **coding** scheme
•To add sparsity constraints

Codings of words is restricted by **group lasso**
•To align the sparse pattern of words' topics

Word count is generated from **Poisson**
•To recover the document 's topic proportion from code

3. MODEL DETAIL

Graphic model of GSTC



Generating process of a document

- For each topic $k \in \{1,\dots,K\}$, sample a word code vector $s_{\cdot k} \in R^N$ ~M-Laplace(λ)
- For each observed word $n \in I$
  - For each topic $k \in \{1,\dots,K\}$, sample a latent word count w_{nk} ~Poisson($s_{nk}\beta_{kn}$)
- Obtain the word count $w_n = \sum_{k=1}^{K} w_{nk}$

Object function

$$\min_{\Theta,\beta} \mathcal{L}(\Theta,\beta) = -\ln P(\Theta,\beta \mid D)$$
$$= \min_{\Theta,\beta} \sum_{d=1}^{M}\sum_{n=1}^{I_d}\left(\sum_{k=1}^{K} s_{dnk}\beta_{kn} - w_{dn}\ln\left(\sum_{k=1}^{K} s_{dnk}\beta_{kn}\right)\right)$$
$$+ \sum_{d=1}^{M}\sum_{k=1}^{K}\lambda \| s_{d\cdot k}\| + C$$

$$s.t.\quad \mathbf{s}_{d,n\cdot} \geq 0, \forall d, n \in \mathbf{I}_d, \sum_{n=1}^{N}\beta_{kn} = 1, \forall k$$

Parameter estimation

1. Fix β, learn $s_{d\cdot}$ for each document d.(block coordinate descent are used for group sparsity)
2. Fix s, learn β with projected gradient method
3. Go to step 1 until converge

From coding to topics

Let θ be the topic proportion vector of document d, then θ_k can be obtained by(formula 2 is derived by the Moran's Property):

$$\theta_k = \mathbb{E}\left[\frac{\sum_{n=1}^{|I|} w_{nk}}{\sum_{n=1}^{|I|}\sum_{k=1}^{K} w_{nk}} \,\middle|\, \sum_{n=1}^{|I|}\sum_{k=1}^{K} w_{nk}\right] = \frac{\mathbb{E}[\sum_{n=1}^{|I|} w_{nk} \mid \sum_{n=1}^{|I|}\sum_{k=1}^{K} w_{nk}]}{\sum_{n=1}^{|I|} w_n} = \frac{\sum_{n=1}^{|I|} \mathbf{s}_{nk}\boldsymbol{\beta}_{kn}}{\sum_{n=1}^{|I|}\sum_{k=1}^{K} \mathbf{s}_{nk}\boldsymbol{\beta}_{kn}} \quad (1)$$

$$\mathbb{E}[\sum_{n=1}^{|I|} w_{nk} \mid \sum_{n=1}^{|I|}\sum_{k=1}^{K} w_{nk}] = \left(\sum_{n=1}^{|I|} w_n\right)\left(\frac{\sum_{n=1}^{|I|} s_{nk}\beta_{kn}}{\sum_{n=1}^{|I|}\sum_{k=1}^{K} s_{nk}\beta_{kn}}\right) \quad (2)$$

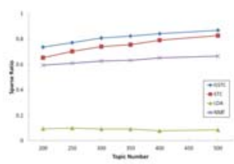
Moran's Property of Poisson Distribution

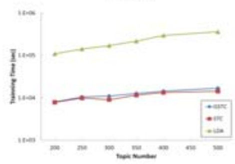
Let $x_1,\dots,x_n$ are independent Poisson random variables with parameters $\tau_1,\dots,\tau_n$, then

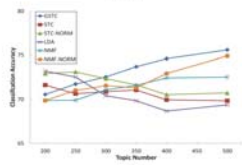
$$x_i \mid \sum_{j=1}^{n} x_j \sim Binom\left(\sum_{j=1}^{n} x_j, \frac{\tau_i}{\sum_{j=1}^{n}\tau_j}\right)$$

4. EXPERIMENTAL RESULTS

- Dataset
  - 20-newsgoup
    - 18, 846 document,
    - 26, 214 distinct words
    - 20 related categories
- Baseline methods
  - LDA, NMF, STC
- Evaluation
  - Topic sparsity
  - Train time
  - Accuracy of document classification







5. CONCLUSIONS

Conclusion

- GSTC provides an elegant way to model topics concerning both sparsity and semantic representation.
- Experiments show the good performance of GSTC in meaningful compact latent representations and document classification.

Future work

- Consider the sparsity of dictionary.
- Develop a paralleled algorithm for large-scale applications.
- Extend the GSTC by integrating the discriminative features of document.