Hierarchically Supervised Latent Dirichlet Allocation

A. Perotte[1] N. Bartlett[2] N. Elhadad[1] F. Wood[2]

[1]Department of Biomedical Informatics
Columbia University

[2]Department of Statistics
Columbia University

NYAS Annual Machine Learning Symposium, 2011

Introduction

- HSLDA: Hierarchically Supervised Latent Dirichlet Allocation
- Model of documents and labels
  - Structure in label space
- Large, real-world datasets

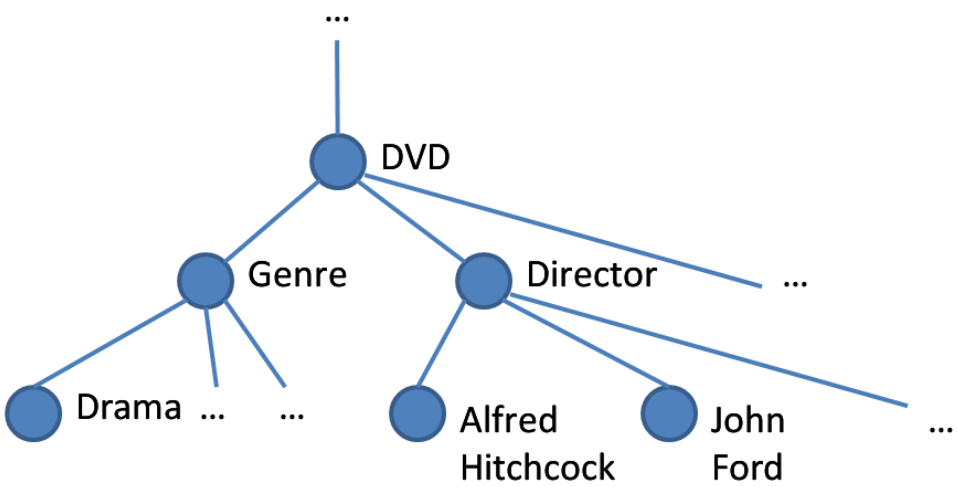
Introduction

- HSLDA: Hierarchically Supervised Latent Dirichlet Allocation
- Model of documents and labels
  - Structure in label space
- Large, real-world datasets

Introduction

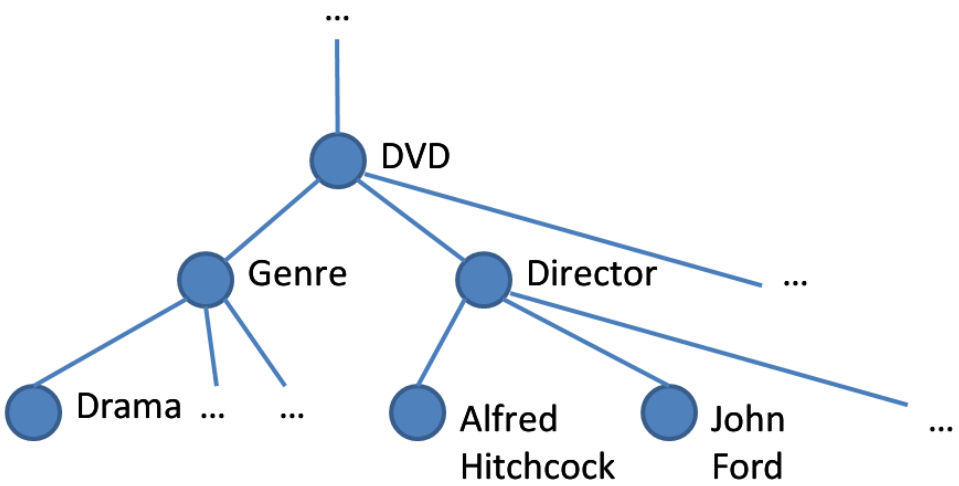
- HSLDA: Hierarchically Supervised Latent Dirichlet Allocation
- Model of documents and labels
  - Structure in label space
- Large, real-world datasets

Amazon.com Data



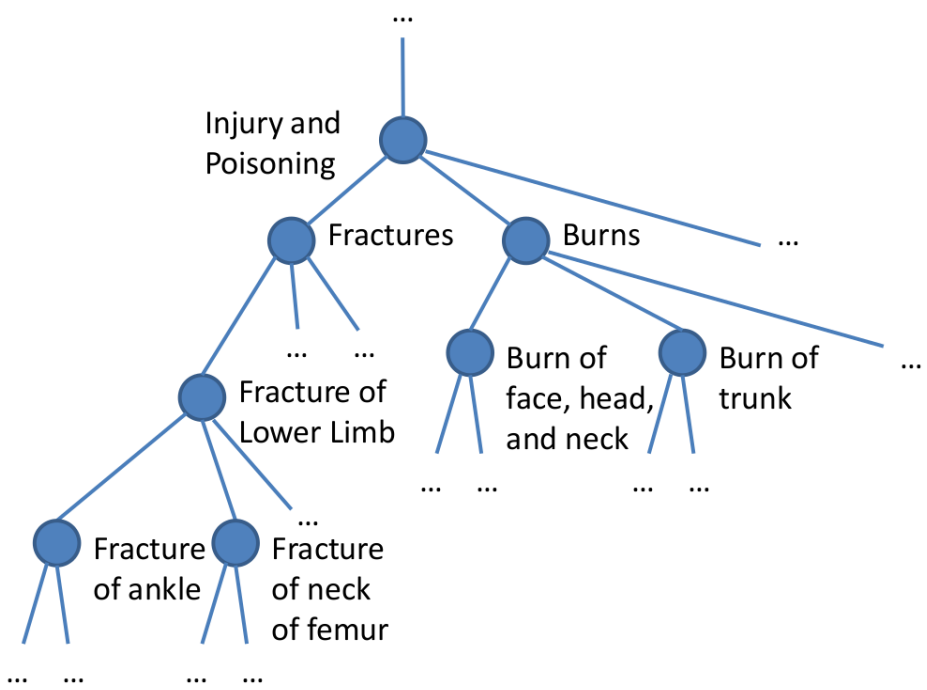
- Text: Product Descriptions: ∼90 words/document
- Labels: Product Categories: ∼9 categories/document

Amazon.com Data



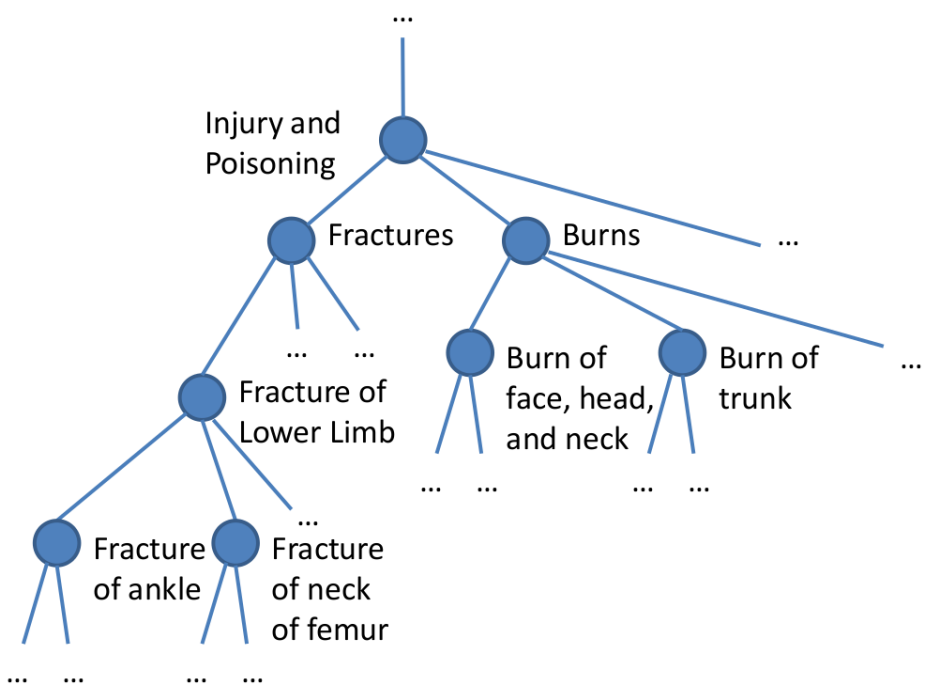
- Text: Product Descriptions: ∼90 words/document
- Labels: Product Categories: ∼9 categories/document

Clinical Data



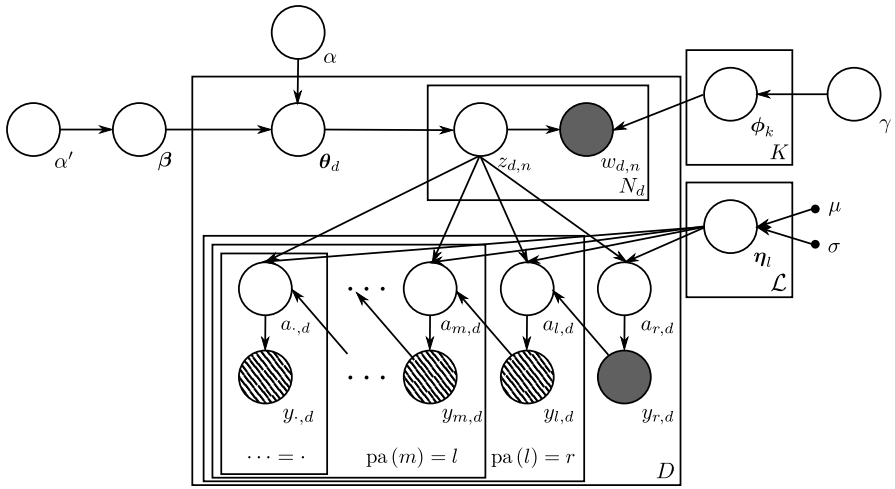
- Text: Discharge summaries: ∼500 words/document
- Labels: ICD9 codes: ∼8 codes/document

Clinical Data



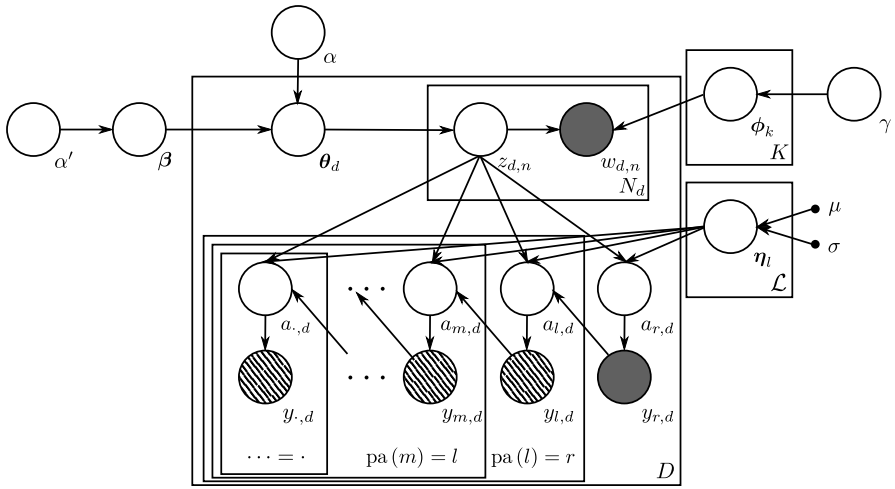
- Text: Discharge summaries: ∼500 words/document
- Labels: ICD9 codes: ∼8 codes/document

Model



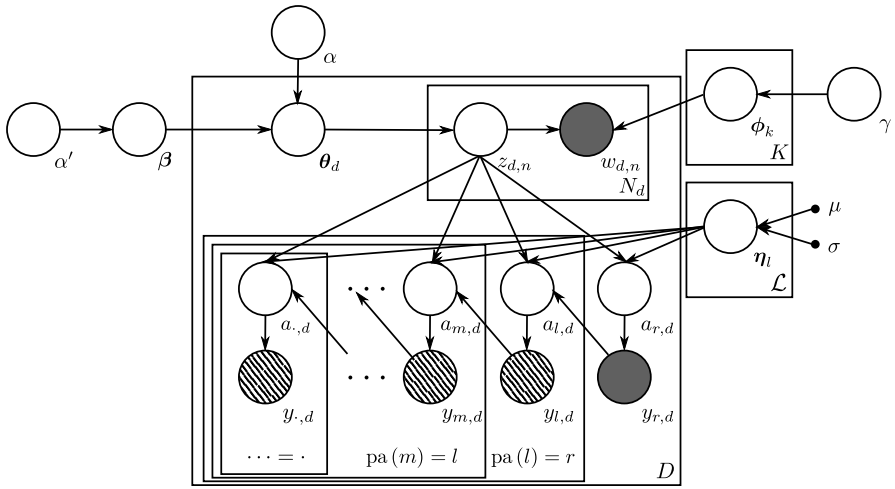
- Documents have latent structure
  - Points in low-dimensional space
- Latent dimensions
  - Distribution over words
- Regression parameters
  - Relationship between the latent space and the label space

Model



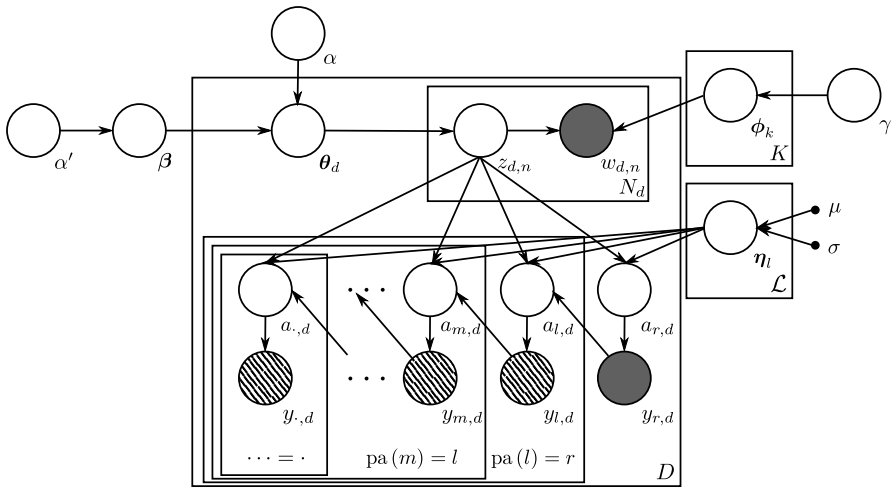
- Documents have latent structure
  - Points in low-dimensional space
- Latent dimensions
  - Distribution over words
- Regression parameters
  - Relationship between the latent space and the label space

Model



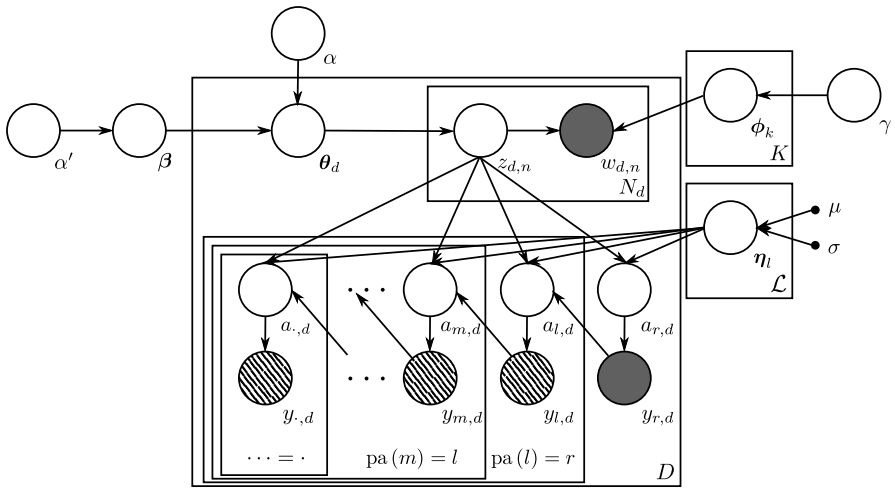
- Documents have latent structure
  - Points in low-dimensional space
- Latent dimensions
  - Distribution over words
- Regression parameters
  - Relationship between the latent space and the label space

Inference



- Collapsed Gibbs sampler
- Probit regression
  - Auxiliary variables allow for Gibbs sampling

Inference



- Collapsed Gibbs sampler
- Probit regression
  - Auxiliary variables allow for Gibbs sampling

Experiments

- Comparison models
  - sLDA with independent regressors
  - HSLDA fit by first performing LDA then fitting tree-conditional regressions
- Task: Prediction of out of sample labels

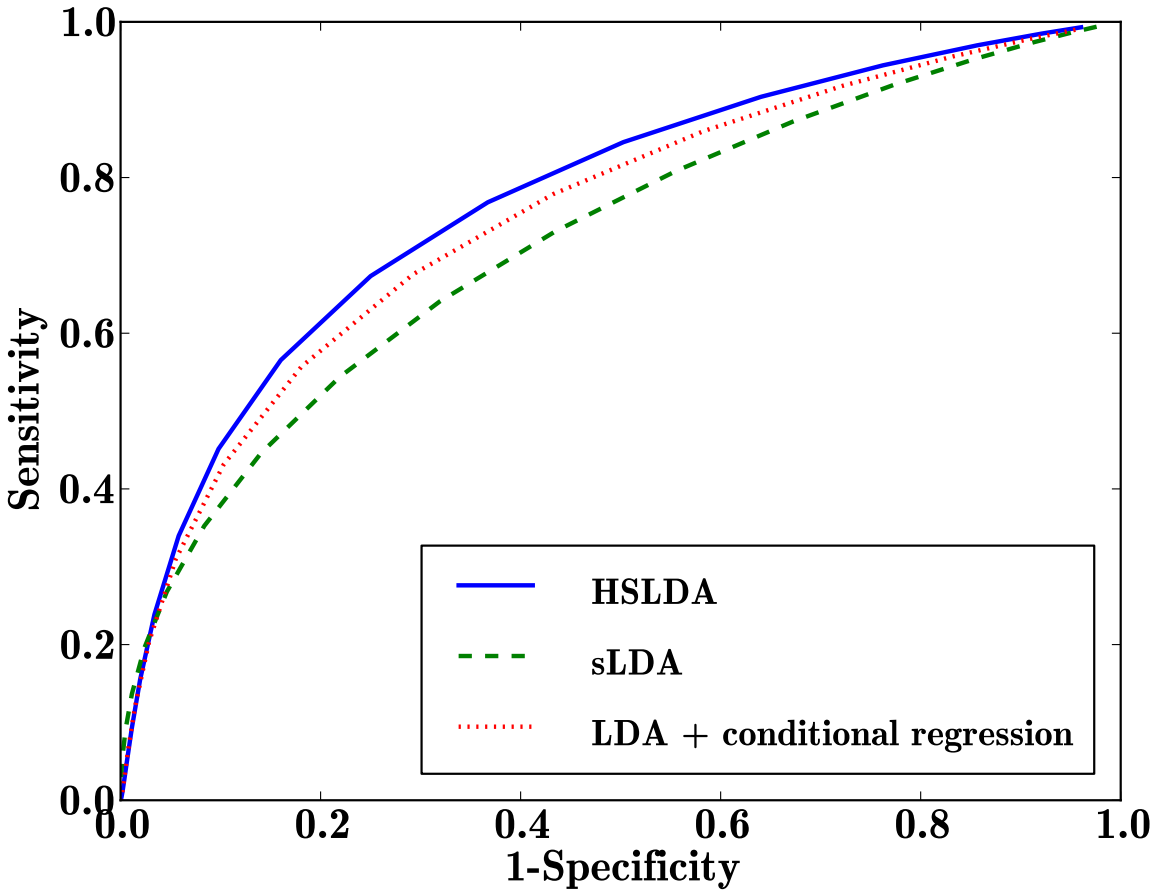
Experiments

- Comparison models
  - sLDA with independent regressors
  - HSLDA fit by first performing LDA then fitting tree-conditional regressions
- Task: Prediction of out of sample labels

Example Topics

Clinical Topics		Product Topics	
MASS	WOUND	SERIES	BASEBALL
CANCER	FOOT	EPISODES	TEAM
RIGHT	CELLULITIS	SHOW	GAME
BREAST	ULCER	SEASON	PLAYERS
CHEMOTHERAPY	LEFT	EPISODE	BASKETBALL
METASTATIC	ERYTHEMA	FIRST	SPORT
LEFT	PAIN	TELEVISION	SPORTS
LYMPH	SWELLING	SET	NEW
TUMOR	SKIN	TIME	PLAYER
BIOPSY	RIGHT	TWO	SEASON
CARCINOMA	ABSCESS	SECOND	LEAGUE
LUNG	LEG	ONE	FOOTBALL
CHEMO	OSTEOMYELITIS	CHARACTERS	STARS
ADENOCARCINOMA	TOE	DISC	FANS
NODE	DRAINAGE	GUEST	FIELD

Prediction


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
HSLDA
sLDA
LDA + conditional regression

Summary

- HSLDA is a new topic model based on sLDA with hierarchical supervision.
- We derive an efficient Gibbs sampler for HSLDA.
- Label prediction can be improved with HSLDA if there exists significant structure in the label space.

Thank you!

- George Hripcsak, MD, MS
- New York Academy of Sciences
- National Library of Medicine