

# A Latent Variable Model for Geographic Lexical Variation

Nicholas Gow

November 20, 2015

# Introduction

- ▶ Different from standard dialectometry
- ▶ “Assumption free”
  - ▶ Unsupervised model for discovering dialect regions
  - ▶ Unsupervised model for discovering latent topics in the text
  - ▶ Supervised model for geolocation prediction
- ▶ Large dataset
  - ▶ Noisy Twitter data

# What are latent variable models?

Also known as Graphical models, Probabilistic models, Hierarchical models

## *Generative Story*

A probabilistic formulation of the process that generated the data

## Bayesian inference

- ▶ Give the model some data
- ▶ The algorithm will find try to find good parameters
  - ▶ Until convergence
- ▶ Bayes Rule:  $P(\theta_i|D) \propto P(D|\theta_i)P(\theta)$
- ▶  $P(D|\theta_i)$  is the likelihood of the data given the parameters
- ▶  $P(\theta)$  is the *prior* probability of the parameters

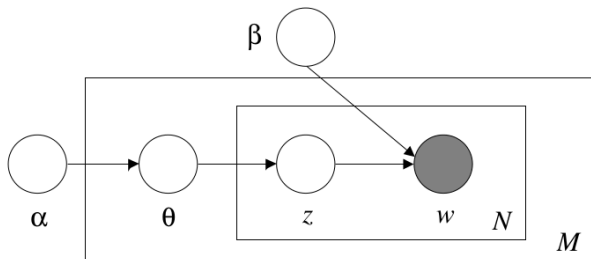
# Latent Dirichlet Allocation (LDA) [1]

- ▶ Find latent topics in a corpus
  - ▶ For any document, tell me which topics are in it
  - ▶ For any word, tell me what topics it belongs to
  - ▶ For any topic, tell me which words belong to it

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

# LDA: Generative story

1. Choose  $N \sim \text{Poisson}(\xi)$ .
2. Choose  $\theta \sim \text{Dir}(\alpha)$ .
3. For each of the  $N$  words  $w_n$ :
  - (a) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$ .
  - (b) Choose a word  $w_n$  from  $p(w_n | z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$ .



# Caveats about latent variable models

- ▶ Independence assumptions
  - ▶ Topics are chosen independently
  - ▶ Bag of words
- ▶ Tractability (models cannot be arbitrarily complex)

# Model: Geographic Topic Model [2]

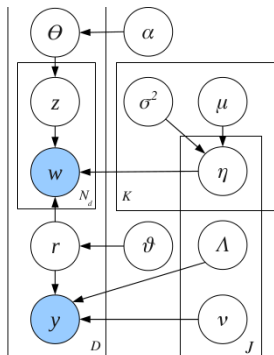
- **Generate base topics:** for each topic  $k < K$ 
  - Draw the base topic from a normal distribution with uniform diagonal covariance:  $\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{a}, b^2 \mathbf{I})$ ,
  - Draw the regional variance from a Gamma distribution:  $\sigma_k^2 \sim \mathcal{G}(c, d)$ .
  - **Generate regional variants:** for each region  $j < J$ ,
    - \* Draw the region-topic  $\boldsymbol{\eta}_{jk}$  from a normal distribution with uniform diagonal covariance:  $\boldsymbol{\eta}_{jk} \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 \mathbf{I})$ .
    - \* Convert  $\boldsymbol{\eta}_{jk}$  into a multinomial distribution over words by exponentiating and normalizing:  
$$\boldsymbol{\beta}_{jk} = \exp(\boldsymbol{\eta}_{jk}) / \sum_i^W \exp(\eta_{jk}^{(i)})$$
where the denominator sums over the vocabulary.

# Model: Geographic Topic Model - continued

- **Generate regions:** for each region  $j < J$ ,
  - Draw the spatial mean  $\boldsymbol{\nu}_j$  from a normal distribution.
  - Draw the precision matrix  $\Lambda_j$  from a Wishart distribution.
- Draw the distribution over regions  $\boldsymbol{\vartheta}$  from a symmetric Dirichlet prior,  $\boldsymbol{\vartheta} \sim \text{Dir}(\alpha\mathbf{1})$ .
- **Generate text and locations:** for each document  $d$ ,
  - Draw topic proportions from a symmetric Dirichlet prior,  $\boldsymbol{\theta} \sim \text{Dir}(\alpha\mathbf{1})$ .
  - Draw the region  $r$  from the multinomial distribution  $\boldsymbol{\vartheta}$ .
  - Draw the location  $\mathbf{y}$  from the bivariate Gaussian,  $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\nu}_r, \Lambda_r)$ .
  - For each word token,
    - \* Draw the topic indicator  $z \sim \boldsymbol{\theta}$ .
    - \* Draw the word token  $w \sim \beta_{rz}$ .



# Model: Plate Notation



$\mu_k$	log of base topic $k$ 's distribution over word types
$\sigma_k^2$	variance parameter for regional variants of topic $k$
$\eta_{jk}$	region $j$ 's variant of base topic $\mu_k$
$\theta_d$	author $d$ 's topic proportions
$r_d$	author $d$ 's latent region
$y_d$	author $d$ 's observed GPS location
$\nu_j$	region $j$ 's spatial center
$\Lambda_j$	region $j$ 's spatial precision
$z_n$	token $n$ 's topic assignment
$w_n$	token $n$ 's observed word type
$\alpha$	global prior over author-topic proportions
$\vartheta$	global prior over region classes

How inference is run: *Variational Inference*[3] (won't discuss here)

# Model: Data

- ▶ 15% of all public messages over one week
- ▶ Collect geo-tagged data from Twitter
- ▶ Filter data
  - ▶ Only authors who wrote at least 20 messages
  - ▶ Only authors follow fewer than 1000 people and are followed by fewer than 1000
  - ▶ No messages containing urls
  - ▶ Only the 49 contiguous US States
- ▶ 9,500 users
- ▶ 380,000 messages
- ▶ 4.7 million tokens
- ▶ Developed useful Twitter tokenizer as part of this work [4]

## Model: Initialization

- ▶ Dirichlet Mixture model (clustering algorithm which finds a suitable number of clusters) to find number of regions ( $J$ )
- ▶ Several k-means runs to get an idea about within region variance (dispersion)
- ▶ LDA run to get estimates of  $z$  for each token

# Evaluation: metric for geolocation prediction

## Regression

- ▶ Mean/Median distance from true location
- ▶ Median is probably better because of potential for outliers

## Classification

- ▶ 4-way (regional) Random classifier - 25%
- ▶ 49-way (state) Random classifier - 2%

# Evaluation: Competing models

## Regions only (without topics $\rightarrow K=1$ )

- ▶ Can set up the geographic topic model to have only one topic
- ▶ The performance of this model relative to the geographic topic model tests the assumption that jointly modelling topics is useful for the regional prediction task

## Text Regression

- ▶ TFIDF vectors
- ▶ train *coupled* regressions for latitude and longitude

## K-means

Should work better than text regression, because it allows for complex partitionings of the output space.

# Evaluation: Competing models: Supervised LDA [5]

- ▶ A variant of LDA that jointly models the data together with the output values/labels

1. Draw topic proportions  $\theta | \alpha \sim \text{Dir}(\alpha)$ .
2. For each word
  - (a) Draw topic assignment  $z_n | \theta \sim \text{Mult}(\theta)$ .
  - (b) Draw word  $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$ .
3. Draw response variable  $y | z_{1:N}, \eta, \delta \sim \text{GLM}(\bar{z}, \eta, \delta)$ , where we define

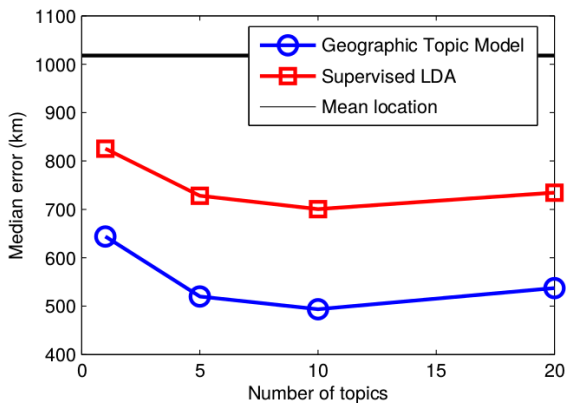
$$(1) \quad \bar{z} := (1/N) \sum_{n=1}^N z_n.$$

Figure : Supervised LDA *generative story*

# Evaluation: Regression/Classification Results

System	Regression		Classification accuracy (%)	
	Mean Dist. (km)	Median Dist. (km)	Region (4-way)	State (49-way)
Geographic topic model	<b>900</b>	<b>494</b>	<b>58</b>	24
Mixture of unigrams	947	644	53	19
Supervised LDA	1055	728	39	4
Text regression	948	712	41	4
<i>K</i> -nearest neighbors	1077	853	37	2
Mean location	1148	1018		
Most common class			37	<b>27</b>

## Evaluation: Accounting for topics helps!

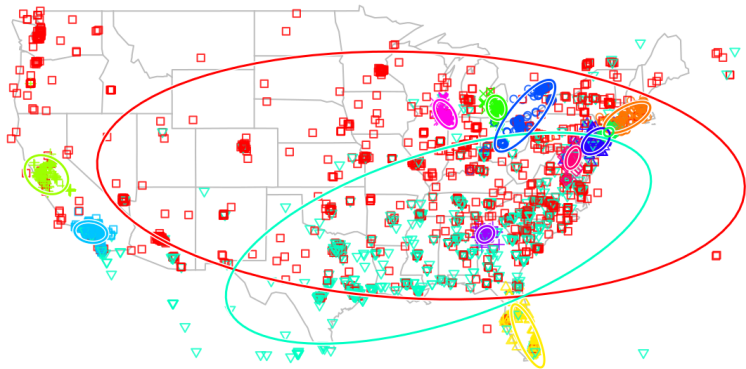




# Exploratory data analysis: Topics discovered

	“basketball”	“popular music”	“daily life”	“emoticons”	“chit chat”
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :( :) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	:p gna loveee	<i>ese</i> exam suttin sippin
N. California 	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	<i>pues</i> hella koo SAN fckn	hella flirt hut iono OAKLAND
New York 	NETS KNICKS	BRONX	iam cab	oww	wassup nm
Los Angeles 	#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af <i>papi</i> raining th bomb coo HOLLYWOOD	wyd coo af <i>nada</i> tacos messin fasho bomb
Lake Erie 	CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	:d blvd BIEBER hve OHIO	foul WIZ salty excuses lames officer lastnight






# Exploratory data analysis: Regions discovered



# Conclusion

- ▶ A lot of parameters in the model - was there enough data
- ▶ A lot of noise in the input
- ▶ A lot of noise in the output
- ▶ Variational Inference for Geographic Topic Model non-trivial to implement

# Bibliography

-  David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *the Journal of machine Learning research 3* (2003), pp. 993–1022.
-  Jacob Eisenstein et al. “A latent variable model for geographic lexical variation”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2010, pp. 1277–1287.
-  Martin J Wainwright and Michael I Jordan. “Graphical models, exponential families, and variational inference”. In: *Foundations and Trends® in Machine Learning 1.1-2* (2008), pp. 1–305.
-  Kevin Gimpel et al. “Part-of-speech tagging for twitter: Annotation, features, and experiments”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics. 2011, pp. 42–47.
-  David M Blei and Jon D McAuliffe. “Supervised Topic Models”. In: *arXiv preprint arXiv:1003.0783* (2010).

# Discussion