Bayesian Nonparametric Topic Modeling for Large Spatial Data

Julian Straub and John W. Fisher III
Massachusetts Institute of Technology (MIT)

Aim: Predictions based on location type and not on location

$p(\text{model}|\text{location}) \rightarrow p(\text{model}|\text{topic distribution})$

The car dataset

- 696k sensor time-associated measurements from 1k trips with a standard car
- trips are all in the greater Boston area

Hierarchical Dirichlet Process (HDP) model of a car's signals

- map ↔ corpus of documents
- road-states ↔ documents
- quantized car measurements ↔ words

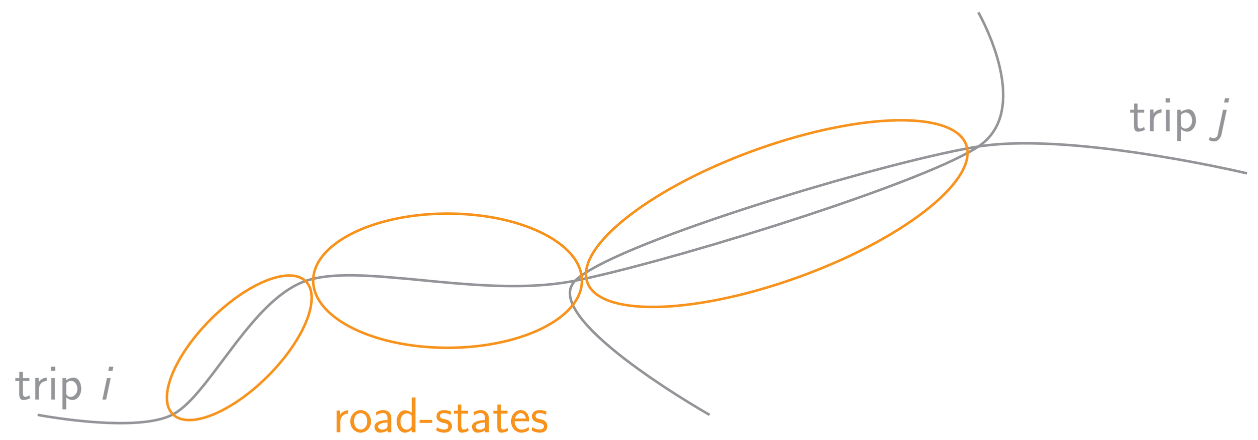
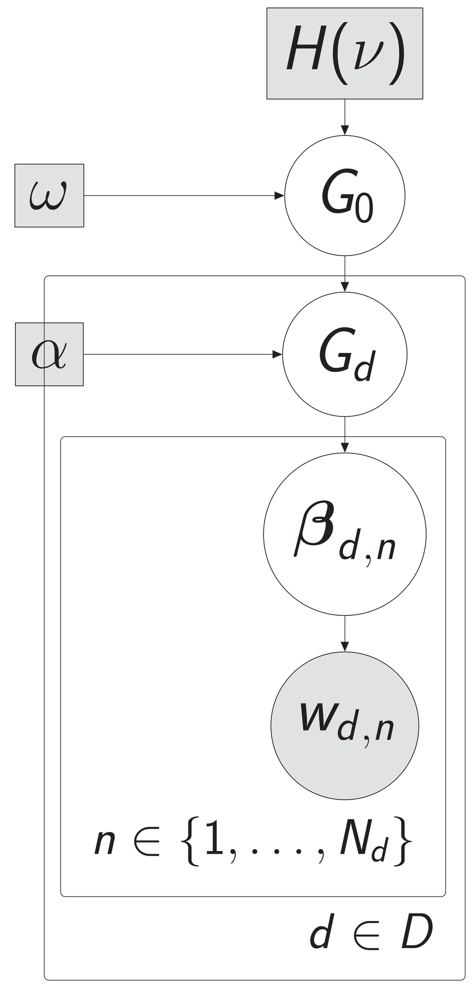


Figure: A road-state is a small segment of a road that is associated with the respective segments of all trips leading trough it.

The gain from the HDP modeling

- hierarchy allows sharing of data across road-states
- characterize driving behavior
- describe driving situation in a road-state

HDP model by [TJBB06]



Notation

- D road-states (=documents).
- N_d observations (=words) per road-state.
- $H(\nu)$ Dirichlet base distribution
- G_0 map (=corpus) level DP
- G_d road-state (=document) level DPs
- $\beta_{d,n}$ Multinomial distribution over $w_{d,n}$
- $w_{d,n}$ quantized measurements (=words)

Gibbs Sampling [TJBB06]

- multiple passes through the whole data
- converges to true distribution (eventually)

Stochastic Variational Inference [WPB11]

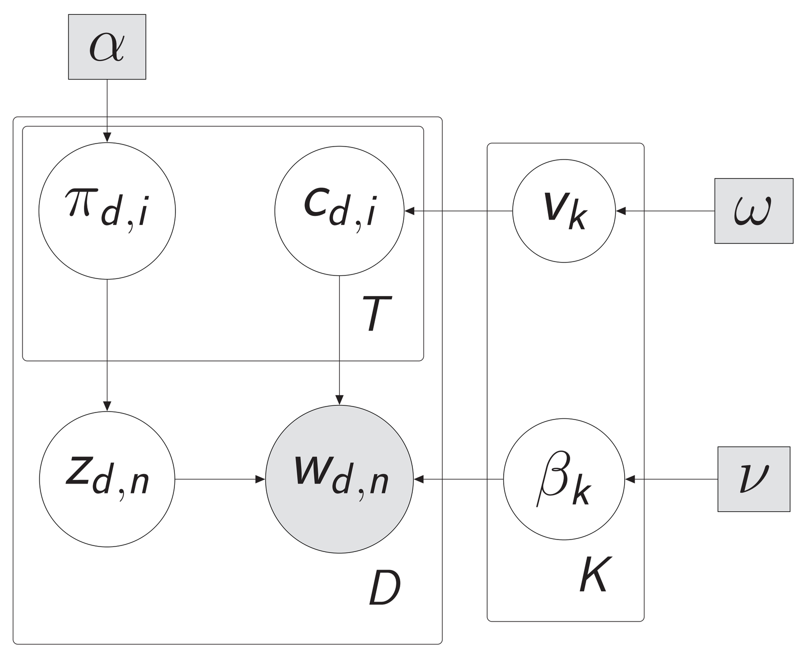
- online version needs only single pass through data
- can be parallelized
- only approximation of true distribution

Comparison Stochastic Variational vs Gibbs Sampling on artificial data

Dataset: $D = 50$ documents with $N_d = 100$ words each; 50 different words.

	Stochastic Variational	Gibbs Sampling
total time	140 s	10,000 s
time per iteration over all D documents	140 s	100 s

HDP model for Stochastic Variational Inference by [WPB11]



Notation

- K corpus level topics
- T document level topics
- D documents
- v_k topic proportions on corpus level
- $c_{d,i}$ indicator selecting corpus level topic
- $\pi_{d,i}$ topic proportions on document level
- $z_{d,n}$ indicator selecting document level topic
- β_k Multinomial distribution over words
- $w_{d,n}$ words $w_{d,n} \sim \text{MULT}\left(\beta_{c_{d,z_{d,n}}}\right)$

References

- Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, *Hierarchical dirichlet processes*, Journal of the American Statistical Association (JASA) **101** (2006), no. 476, 1566–1581.
- Chong Wang, John Paisley, and David M Blei, *Online variational inference for the hierarchical dirichlet process*, Artificial Intelligence and Statistics, 2011.

Results

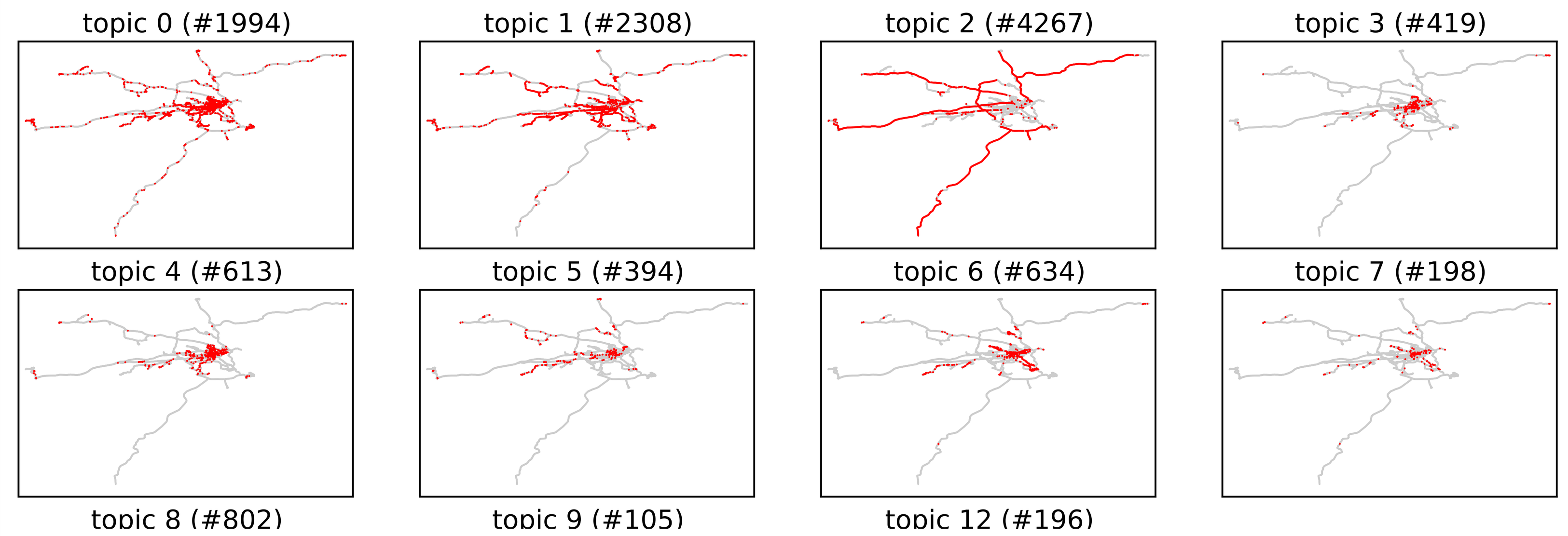


Figure: Road-states split according to their most probable topic

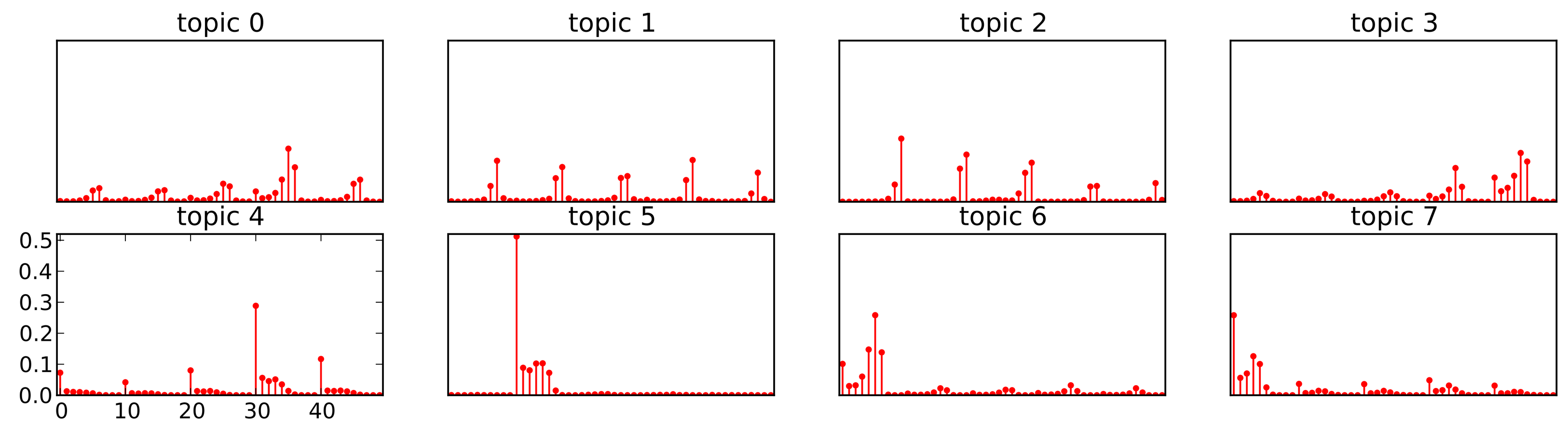


Figure: Word (joint speed and time-of-day) distributions for the topics

Pooling of data across road-states

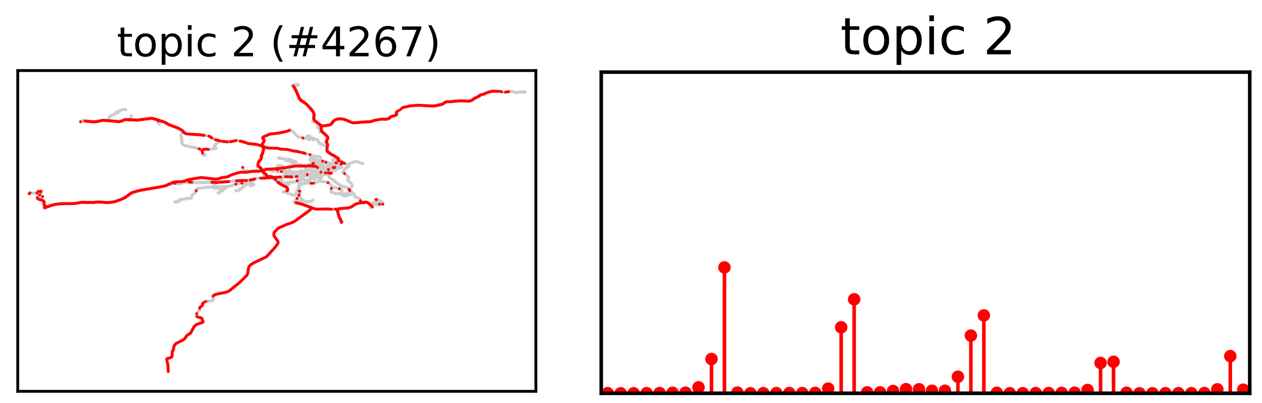


Figure: Note that the road-states used to learn topic 2 contain very few (∼ 10 per state) measurements.

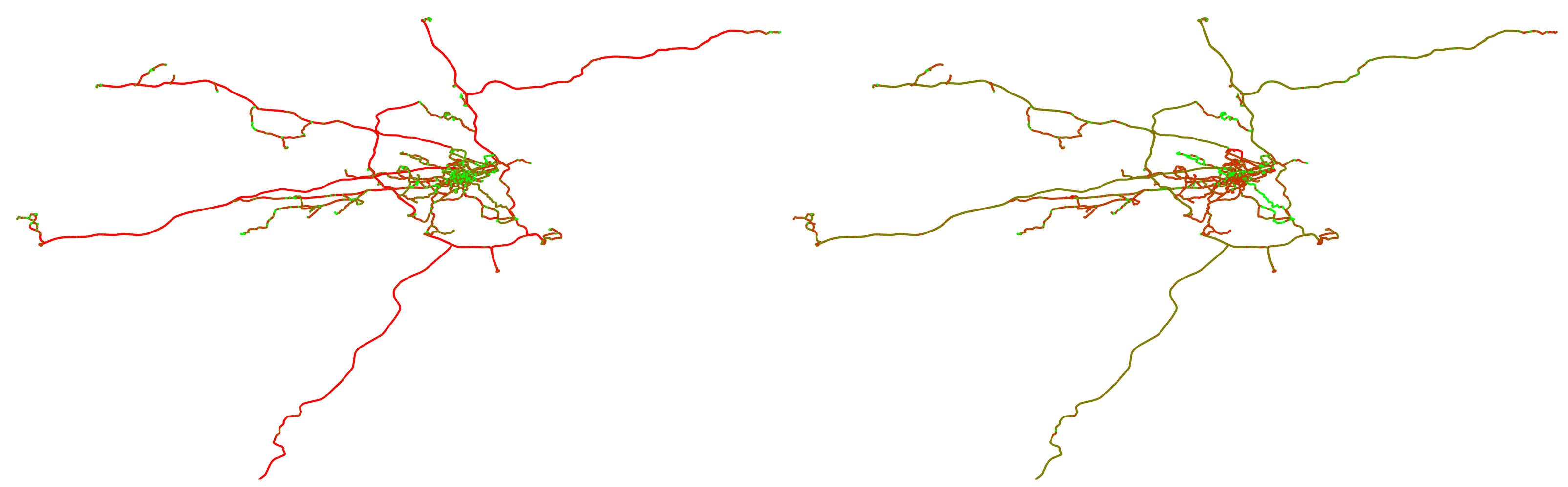


Figure: Marginal distributions on speed (left) and time-of-day (right)

Open Questions

- What to do with a topic model that describes the data?
- How do we compare topic models accros states?
- Bag of Words assumption neglects dependencies - other approaches?

mail: jstraub@csail.mit.edu, fisher@csail.mit.edu www: http://groups.csail.mit.edu/vision/sli/