Non-parametric Methods for Unsupervised Semantic Modelling

Wray Buntine

Monash University

Thanks Lan Du of Macquarie University and Swapnil Mishra and Kar Wai Lim of ANU for many great slides, and Lancelot James for teaching me general IBP theory.

Tuesday 27th January, 2015

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Information Overload

 langwitches' photostream @ Flickr

We need new tools to help us: organize, search, summarise and understand information. The field of Information Access serves this purpose.

Information Warfare

Definition: "the use and management of information in pursuit of a competitive advantage over an opponent."

- Email spam, link spam, *etc.*
  - Whole websites are fabricated with fake content to trick search engines.
  - Spammers using social networks to personalise attacks (Nov. 2011).
- BBC reports trust in information on the web is being damaged "by the huge numbers of people paid by companies to post comments" (Dec. 2011).

It's an information war out there on the internet between consumers (*i.e.*, you), companies, not-for-profits, voters, parties, employees, bureaucrats, academics, ...

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Probability Vectors

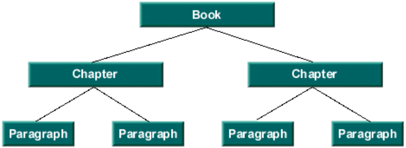
We often have probability vectors for:

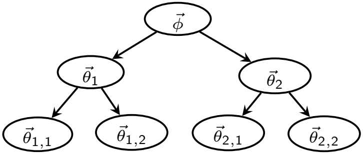
- the next word given $(n-1)$ previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,
- hashtag in a tweet given the author.

We need to work with distributions over probability vectors for:

- inheritance and sharing of information;
- networks of probability vectors;
- inference and learning.

Sharing/Inheritance with a Probability Hierarchy





We might model a set of vocabularies/documents hierarchically:

$$\vec{\theta}_1 \sim \text{Dirichlet}\left(\alpha_0 \vec{\phi}\right)$$
$$\vec{\theta}_{1,2} \sim \text{Dirichlet}\left(\alpha_1 \vec{\theta}_1\right)$$

Statistical estimation with these generally difficult:

$$\ldots \frac{1}{\text{Beta}\left(\alpha_1 \vec{\theta}_1\right)} \prod_k \theta_{1,2,k}^{\alpha_1 \theta_{1,k}-1} \prod_k {\theta_{1,k}}^{\alpha_1 \phi_k - 1} \ldots$$

(**N.B.** fixed point MAP solutions exist for hierarchies)

Overview: Latent Semantic Modelling

- Variety of component and network models can be made non-parametric with deep probability vector networks.

Overview: Latent Semantic Modelling

- Variety of component and network models can be made non-parametric with deep probability vector networks.
- Newer fast methods for training deep probability vector networks:

Overview: Latent Semantic Modelling

- Variety of component and network models can be made non-parametric with deep probability vector networks.
- Newer fast methods for training deep probability vector networks:
  - Chinese Restaurant processes for Pitman-Yor and Dirichlet Processes seems mostly poor;
  - stick-breaking appears to interact badly with variational methods;
  - the minimum path assumption can be poor;
  - concentration parameter often should be fit.

Overview: Latent Semantic Modelling

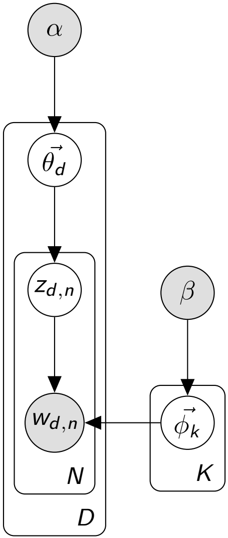
- Variety of component and network models can be made non-parametric with deep probability vector networks.
- Newer fast methods for training deep probability vector networks:
  - Chinese Restaurant processes for Pitman-Yor and Dirichlet Processes seems mostly poor;
  - stick-breaking appears to interact badly with variational methods;
  - the minimum path assumption can be poor;
  - concentration parameter often should be fit.
- Allows efficient modelling of latent semantics:
  - semantic resources to integrate (WordNet, sentiment dictionaries, *etc.*),
  - inheritance and shared learning across multiple instances,
  - hierarchical modelling,
  - deep latent semantics,
  - integrating semi-structured and networked content,

i.e. Same as deep neural networks!

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Latent Dirichlet Allocation



LDA Model

$$
\begin{array}{llcl}
\forall_d & \vec{\theta}_d & \sim & \text{Dirichlet}_K(\vec{\alpha}) \\
\forall_k & \vec{\phi}_k & \sim & \text{Dirichlet}_W(\vec{\beta}) \\
\forall_{d,n} & z_{d,n} & \sim & \text{Categorical}(\vec{\theta}_d) \\
\forall_{d,n} & w_{d,n} & \sim & \text{Categorical}(\vec{\phi}_{z_{d,n}})
\end{array}
$$

Collapsed Posterior for Gibbs Sampling

$$
\prod_d \frac{\text{Beta}_K\left(\vec{n}_{\vec{\theta}_d} + \vec{\alpha}\right)}{\text{Beta}_K\left(\vec{\alpha}\right)} \prod_k \frac{\text{Beta}_W\left(\vec{n}_{\vec{\phi}_k} + \vec{\beta}\right)}{\text{Beta}_W\left(\vec{\beta}\right)}
$$

Learning Algorithms with Dirichlets

Common text-book algorithms/methods in modern machine-learning/statistics rely on Dirichlet distributions combined with:

- trees, tables;
- graphs, networks;
- context free grammars;

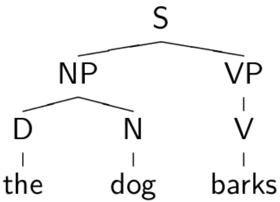
Algorithms on these combine Dirichlet normalizers with:

- model search;
- model averaging;
- EM algorithm;
- *etc.*

Arguably, many of these are of poor/mixed quality.
e.g., probabilistic context free grammars, decision trees

Context Free Grammar

$$\mathcal{R} = \left\{ \begin{array}{ccc} S \rightarrow NP\ VP & NP \rightarrow D\ N & VP \rightarrow V \\ D \rightarrow \text{the} & N \rightarrow \text{dog} & V \rightarrow \text{barks} \end{array} \right\}$$



$$\begin{array}{cc} & S \\ \Rightarrow & NP\ VP \\ \Rightarrow & D\ N\ VP \\ \Rightarrow & \text{the } N\ VP \\ \Rightarrow & \text{the dog } VP \\ \Rightarrow & \text{the dog } V \\ \Rightarrow & \text{the dog barks} \end{array}$$

In a probabilistic context free grammar, probabilities are associated with each rule, and rules apply independently of context.

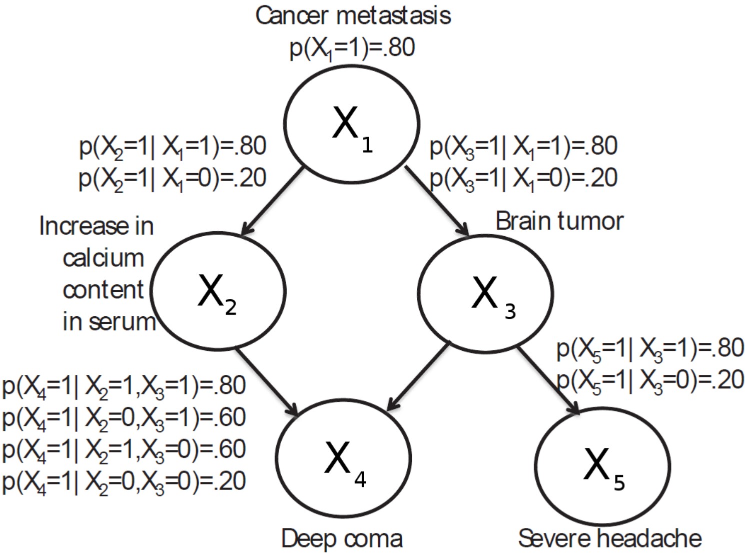
Probabilistic Context Free Grammar, cont.

- Doesn't perform well in practice because statistically, context matters.

  Google bought Youtube.

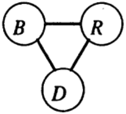
- Previous words, or higher parts-of-speech do affect probabilities.
- State-of-the-art NLP systems "hack" context by making probabilities dependent on:
  - previous few words in the input stream;
  - previous few parts-of-speech higher in the parse tree;
  - head ("main") words for nodes higher in the parse tree.
- Alternatively, they introduce specialised parts-of-special to introduce context.
- This requires algorithmic sophistication!

Bayesian Networks



Bayesian Model Averaging (BMA) for Bayesian Networks

Summaries of the posterior distributions of N for the spina bifida data for all models with posterior probability greater than 0.01. $\hat{N}$ is a Bayes estimate, minimizing a relative squared error loss function

Model	Posterior Prob.	$\hat{N}$	2.5%, 97.5%
	0.373	731	(701, 767)
	0.301	756	(714,811)
	0.281	712	(681,751)
	0.036	697	(628,934)
Model Averaging	—	731	(682,797)

From Madigan and York, 1995

Bayesian Model Averaging (BMA) for Bayesian Networks, cont

Build a pool of "good" models (*i.e.*, the graphical structures) based on the training data: Good-Models($\mathcal{X}$).

BMA estimate of probability for new data $\vec{x}$ given training sample $\mathcal{X}$ is

$$p(\vec{x}|\mathcal{X}) \approx \sum_{M \in \text{Good-Models}(\mathcal{X})} \frac{p(M|\mathcal{X})}{\sum_{M \in \text{Good-Models}(\mathcal{X})} p(M|\mathcal{X})} p(\vec{x}|M,\mathcal{X})$$

Question: how do we build "good" models given there are a combinatoric number?

Bayesian Model Averaging and Non-parametrics Storyline

1990: My PhD thesis, BMA for decision trees.

1990: York and Madigan develop BMA for Bayesian networks.

1994: Breiman developed bagging (or random forests) for trees as a Frequentist response:

→ still one of the top performing classification algorithms

1995: Willems, Shtarkov, Tjalkens adapt BMA for n-grams, context tree weighting (CTW) for lossless compression.

Bayesian model averaging and Frequentist bagging became dominant paradigms.

Bayesian Model Averaging and Non-parametrics Storyline, cont.

2006: Y.W. Teh develops hierarchical Pitman-Yor model for n-grams.

2009: Gasthaus, Wood, Archambeau, Teh and James develop Sequence Memoizer for n-grams for lossless compression. Beats CTW.

2009: Wood and Teh develop Statistical Language Model Domain Adaptation. Further improves n-gram modelling by allowing adaptation.

- but the algorithm is impractical

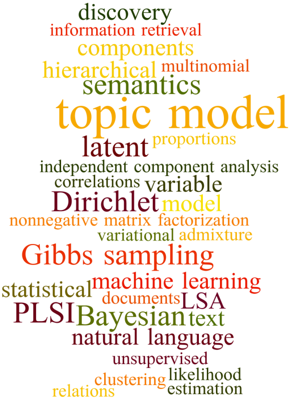
Non-parametric Bayesian methods give new life to BMA because they use **substantially better priors.**

Motivation

- While somewhat successful, the BMA paradigm based on standard (simple) conjugate priors has reached a limit for some models consisting of Dirichlets.
- But this requires efficient non-parametric modelling.
- Which we now have.

Many problems in machine learning are ripe for improvement with better modelling of context.

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Dirichlet Process

The Dirichlet Process (DP) has two arguments $\text{DP}\left(\alpha, H(\cdot)\right)$:

- $H(\cdot)$ is another distribution that is the mean of the DP
- when $H(\cdot) = \vec{\mu}$, applied to a finite probability vector $\vec{\mu}$ of dimension K, the DP and the Dirichlet are identical:

$$\text{Dirichlet}_K\left(\alpha, \vec{\mu}\right) \;=\; \text{DP}\left(\alpha, \vec{\mu}\right) \;.$$

The Pitman-Yor Process (PYP) has three arguments $\text{PYP}\left(d, \alpha, H(\cdot)\right)$:

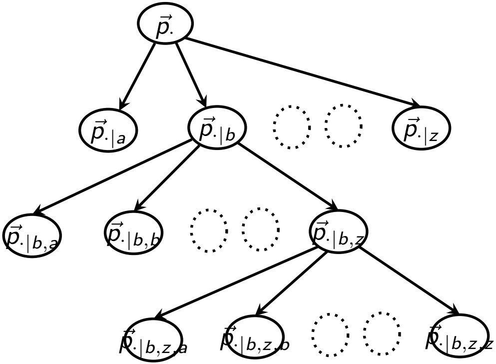
- extends the DP with an extra parameter d, and is more suited to Zipfian data,

Bayesian Idea: Similar Context Means Similar Word

store . Shivering , he caught a ?

- Words in a ? should be like words in ?
  - though no plural nouns
- Words in caught a ? should be like words in a ?
  - though a suitable object for "caught"
- Words in he caught a ? be *very like* words in caught a ?
  - "he" shouldn't change things much

Bayesian N-grams, cont.



$$
\begin{aligned}
\mathcal{S} &= \text{symbol set, fixed or possibly countably infinite} \\
\vec{p}_\cdot &\sim \text{prior on prob. vectors (initial vocabulary)} \\
\vec{p}_{\cdot|x_1} &\sim \text{dist. on prob. vectors with mean } \vec{p}_\cdot \qquad \forall_{x_1 \in \mathcal{S}} \\
\vec{p}_{\cdot|x_1,x_2} &\sim \text{dist. on prob. vectors with mean } \vec{p}_{\cdot|x_1} \qquad \forall_{x_1,x_2 \in \mathcal{S}}
\end{aligned}
$$

Historical Context

1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*

2001-2003: Ishwaran and James develops and "translates" methods usable for machine learning.

2006: Teh develops hierarchical n-gram models using HPYs.

2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes (HDP), *e.g.* applied to LDA.

2006-2011: Chinese restaurant processes (CRPs) go wild!

- require dynamic memory in implementation,
- Chinese restaurant franchise,
- multi-floor Chinese restaurant process,
- *etc.*

Historical Context

1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*

2001-2003: Ishwaran and James develops and "translates" methods usable for machine learning.

2006: Teh develops hierarchical n-gram models using HPYs.

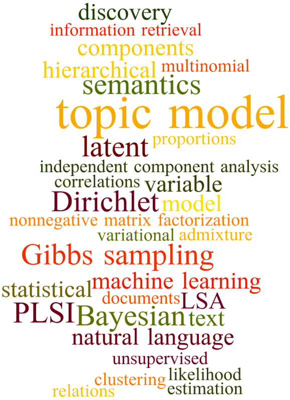
2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes (HDP), *e.g.* applied to LDA.

2006-2011: Chinese restaurant processes (CRPs) go wild!

- require dynamic memory in implementation,
- Chinese restaurant franchise,
- multi-floor Chinese restaurant process,
- *etc.*

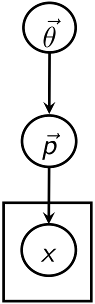
2011: Chen, Du, Buntine show Chinese restaurants and stick-breaking not needed by introducing **block table indicator samplers**.

Outline

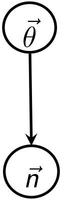

discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

The Ideal Hierarchical Component?

We want a magic distribution that looks like a multinomial likelihood in $\vec{\theta}$.



$$\vec{p} \sim \text{Magic}\left(\alpha, \vec{\theta}\right)$$
$$x_n \sim \text{Discrete}(\vec{p}) \quad \forall_n$$

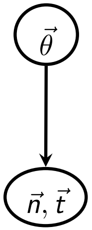


$$p\left(\vec{n} \,\middle|\, \alpha, \vec{\theta}, N\right) = F_\alpha(\vec{n}) \prod_k \theta_k^{t_k}$$

$$\text{where } \sum_k n_k = N$$

The PYP/DP is the Magic

- The PYP/DP plays the role of the magic distribution.
- However, the exponent t_k for the θ now becomes a latent variable, so needs to be sampled as well.
- The t_k are constrainted:
  - $t_k \leq n_k$
  - $t_k > 0$ iff $n_k > 0$
- The $\vec{t}$ act like data for the next level up involving $\vec{\theta}$.



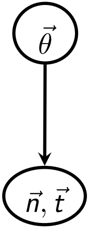
$$p\left(\vec{n}, \vec{t} \,\middle|\, d, \alpha, \vec{\theta}, N\right) = F_{d,\alpha}(\vec{n}, \vec{t}) \prod_k \theta_k^{t_k}$$

where $\sum_k n_k = N$

Interpreting the Auxiliary Counts

Interpretation: t_k is how much of the count n_k that affects the parent probability (*i.e.* $\vec{\theta}$).

- If $\vec{t} = \vec{n}$ then the sample $\vec{n}$ affects $\vec{\theta}$ 100%.
- When $n_k = 0$ then $t_k = 0$, no effect.
- If $t_k = 1$, then the sample of n_k affects $\vec{\theta}$ minimally.



$$p\left(\vec{n}, \vec{t} \,\middle|\, d, \alpha, \vec{\theta}, N\right) = F_{d,\alpha}(\vec{n}, \vec{t}) \prod_k \theta_k^{t_k}$$

where $\sum_k n_k = N$

The Multinomial-Pitman-Yor

Definition of Multinomial-Pitman-Yor

Given a discount d and concentration parameter α, a probability vector $\vec{\theta}$ of dimension L, and a count N, the multinomial-Pitman-Yor creates count vector samples $\vec{n}$ of dimension K, and auxiliary counts $\vec{t}$ (constrained by $\vec{n}$). Now $(\vec{n},\vec{t}) \sim \text{MultPYP}\left(d,\alpha,\vec{\theta},N\right)$ denotes

$$p\left(\vec{n},\vec{t}\,\middle|\,N,\text{MultPYP},d,\alpha,\vec{\theta}\right) = \binom{N}{\vec{n}} \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d}\,\theta_k^{t_k}$$

where $T = \sum_{k=1}^{K} t_k$.

Use rising factorial or Pochhammer symbol $(x|y)_n = x(x+y)...(x+(n-1)y)$, and $(x)_n = (x|1)_n$.

Why We Prefer DPs and PYPs over Dirichlets!

$$p\left(\vec{x} \,\middle|\, N, \text{MultDir}, \alpha, \vec{\theta}\right) \propto$$

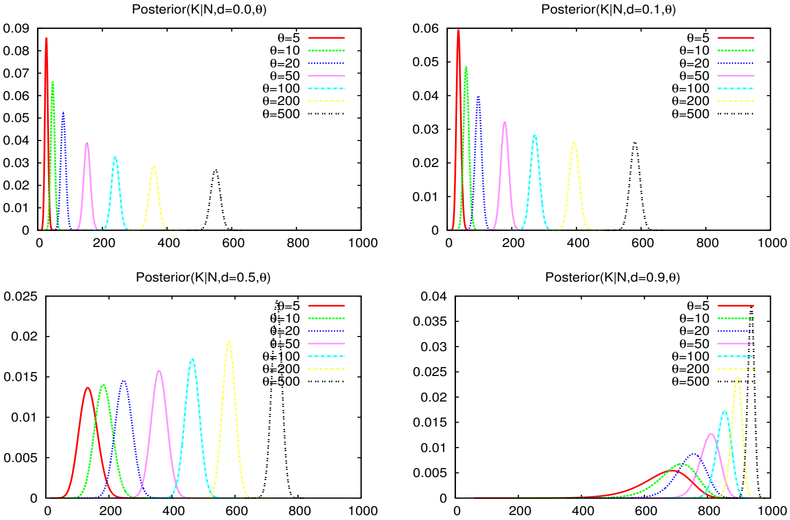
$$\frac{1}{(\alpha)_N} \prod_{k=1}^{K} (\alpha\theta_k)(\alpha\theta_k + 1) \quad \cdots \quad (\alpha\theta_k + n_k - 1)$$

$$p\left(\vec{x}, \vec{t} \,\middle|\, N, \text{MultPYP}, d, \alpha, \vec{\theta}\right) \propto \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} {\theta_k}^{t_k}$$

For the PYP, the θ_k just look like multinomial data, but you have to introduce a discrete latent variable $\vec{t}$.

For the Dirichlet, the θ_k are in a complex gamma function.

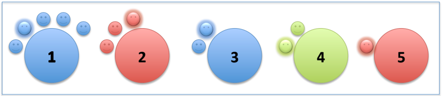
How Many Tables? Why Minimal Path Assumption is Poor



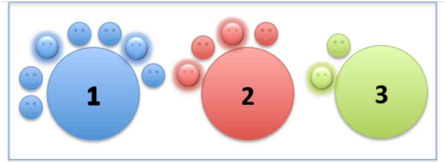
Posterior probability on K given $N = 1000$ and different d, θ.

CRP Samplers versus MultPYP Samplers

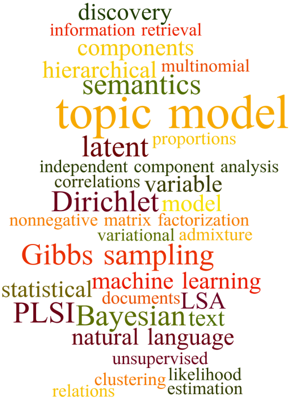
CRP sampling needs to keep track of full seating plan, such as counts per table (thus dynamic memory).



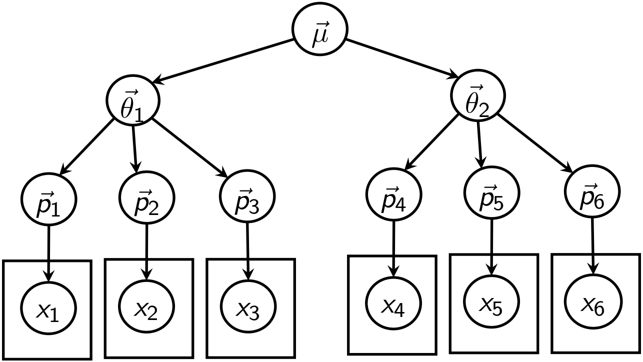
Sampling using the MultPYP formula only needs to keep the number of tables. So rearrange configuration, only one table per dish and mark customers to indicate how many tables the CRP would have had.



Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

A Simple N-gram Style Model



$$p(\vec{\mu})\, p\left(\vec{\theta}_1 \,\middle|\, \vec{\mu}\right) p\left(\vec{\theta}_2 \,\middle|\, \vec{\mu}\right)$$

$$p\left(\vec{p}_1 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_2 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_3 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_4 \,\middle|\, \vec{\theta}_2\right) p\left(\vec{p}_5 \,\middle|\, \vec{\theta}_2\right) p\left(\vec{p}_6 \,\middle|\, \vec{\theta}_2\right)$$

$$\prod_l p_{1,l}^{n_{1,l}} \prod_l p_{2,l}^{n_{2,l}} \prod_l p_{3,l}^{n_{3,l}} \prod_l p_{4,l}^{n_{4,l}} \prod_l p_{5,l}^{n_{5,l}} \prod_l p_{6,l}^{n_{6,l}}$$

Using the Evidence Formula

We will repeatedly apply the evidence formula

$$
\begin{aligned}
p\left(\vec{n}, \vec{t} \,\middle|\, N, \mathrm{MultDP}, \alpha\right) &= \frac{\alpha^{T}}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,0} H(k)^{t_k} \\
&= F_\alpha(\vec{n}, \vec{t}) \prod_{k=1}^{K} H(k)^{t_k}
\end{aligned}
$$

to marginalise out all the probability vectors.

Apply Evidence Formula to Bottom Level

Start with the full posterior:

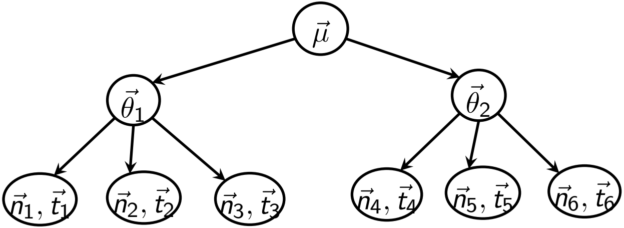
$$p(\vec{\mu})p\left(\vec{\theta}_1 \,\middle|\, \vec{\mu}\right) p\left(\vec{\theta}_2 \,\middle|\, \vec{\mu}\right)$$
$$p\left(\vec{p}_1 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_2 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_3 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_4 \,\middle|\, \vec{\theta}_2\right) p\left(\vec{p}_5 \,\middle|\, \vec{\theta}_2\right) p\left(\vec{p}_6 \,\middle|\, \vec{\theta}_2\right)$$
$$\prod_l p_{1,l}^{n_{1,l}} \prod_l p_{2,l}^{n_{2,l}} \prod_l p_{3,l}^{n_{3,l}} \prod_l p_{4,l}^{n_{4,l}} \prod_l p_{5,l}^{n_{5,l}} \prod_l p_{6,l}^{n_{6,l}} .$$

Marginalise out each $\vec{p}_k$ but introducing new auxiliaries $\vec{t}_k$

$$p(\vec{\mu})p\left(\vec{\theta}_1 \,\middle|\, \vec{\mu}\right) p\left(\vec{\theta}_2 \,\middle|\, \vec{\mu}\right)$$
$$F_\alpha(\vec{n}_1, \vec{t}_1)F_\alpha(\vec{n}_2, \vec{t}_2)F_\alpha(\vec{n}_3, \vec{t}_3) \prod_l \theta_{1,l}^{t_{1,l}+t_{2,l}+t_{3,l}}$$
$$F_\alpha(\vec{n}_4, \vec{t}_4)F_\alpha(\vec{n}_5, \vec{t}_5)F_\alpha(\vec{n}_6, \vec{t}_6) \prod_l \theta_{2,l}^{t_{4,l}+t_{5,l}+t_{6,l}} .$$

Thus $\vec{t}_1 + \vec{t}_2 + \vec{t}_3$ looks like data for $\vec{\theta}_1$ and $\vec{t}_4 + \vec{t}_5 + \vec{t}_6$ looks like data for $\vec{\theta}_2$

Apply Evidence Formula, cont.



Terms left in $\vec{n}_k$ and $\vec{t}_k$, and passing up

$$\prod_l \theta_{1,l}^{t_{1,l}+t_{2,l}+t_{3,l}} \prod_l \theta_{2,l}^{t_{4,l}+t_{5,l}+t_{6,l}} ,$$

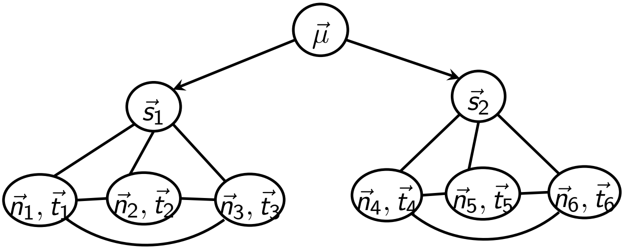
as pseudo-data to the prior on $\vec{\theta}_1$ and $\vec{\theta}_2$.

Apply Evidence Formula, cont.

Repeat the same trick up a level; marginalising out $\vec{\theta}_1$ and $\vec{\theta}_1$ but introducing new auxiliaries $\vec{s}_1$ and $\vec{s}_2$

$$p(\vec{\mu})F_\alpha(\vec{t}_1+\vec{t}_2+\vec{t}_3,\vec{s}_1)F_\alpha(\vec{t}_4+\vec{t}_5+\vec{t}_6,\vec{s}_2)\prod_l \mu_l^{s_{1,l}+s_{2,l}}$$

$$F_\alpha(\vec{n}_1,\vec{t}_1)F_\alpha(\vec{n}_2,\vec{t}_2)F_\alpha(\vec{n}_3,\vec{t}_3)F_\alpha(\vec{n}_4,\vec{t}_4)F_\alpha(\vec{n}_5,\vec{t}_5)F_\alpha(\vec{n}_6,\vec{t}_6)\ .$$



Again left with pseudo-data to the prior on $\vec{\mu}$.

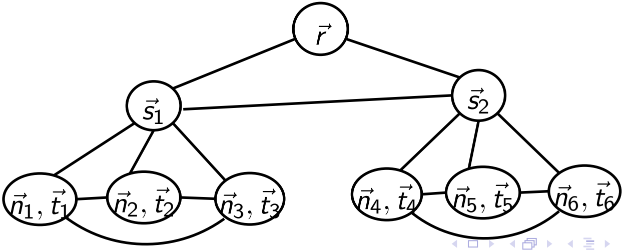
Apply Evidence Formula, cont.

Finally repeat at the top level with new auxiliary $\vec{r}$

$$F_\alpha(\vec{s}_1+\vec{s}_2,\vec{r})F_\alpha(\vec{t}_1+\vec{t}_2+\vec{t}_3,\vec{s}_1)F_\alpha(\vec{t}_4+\vec{t}_5+\vec{t}_6,\vec{s}_2)$$
$$F_\alpha(\vec{n}_1,\vec{t}_1)F_\alpha(\vec{n}_2,\vec{t}_2)F_\alpha(\vec{n}_3,\vec{t}_3)F_\alpha(\vec{n}_4,\vec{t}_4)F_\alpha(\vec{n}_5,\vec{t}_5)F_\alpha(\vec{n}_6,\vec{t}_6)$$

where

- $\vec{n}_1$, $\vec{n}_2$,... are the data at the leaf nodes, $\vec{t}_1$, $\vec{t}_2$,... their auxiliary counts
- $\vec{s}_1$ are auxiliary counts constrained by $\vec{t}_1+\vec{t}_2+\vec{t}_3$,
- $\vec{s}_2$ are auxiliary counts constrained by $\vec{t}_4+\vec{t}_5+\vec{t}_6$,
- $\vec{r}$ are auxiliary counts constrained by $\vec{s}_1+\vec{s}_2$,



The Worked N-gram Style Model

Original posterior in the form:

$$p(\vec{\mu})p\left(\vec{\theta}_1 \,\middle|\, \vec{\mu}\right) p\left(\vec{\theta}_2 \,\middle|\, \vec{\mu}\right)$$

$$p\left(\vec{p}_1 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_2 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_3 \,\middle|\, \vec{\theta}_1\right) p\left(\vec{p}_4 \,\middle|\, \vec{\theta}_2\right) p\left(\vec{p}_5 \,\middle|\, \vec{\theta}_2\right) p\left(\vec{p}_6 \,\middle|\, \vec{\theta}_2\right)$$

$$\prod_l p_{1,l}^{n_{1,l}} \prod_l p_{2,l}^{n_{2,l}} \prod_l p_{3,l}^{n_{3,l}} \prod_l p_{4,l}^{n_{4,l}} \prod_l p_{5,l}^{n_{5,l}} \prod_l p_{6,l}^{n_{6,l}}$$

Collapsed posterior in the form:

$$F_\alpha(\vec{s}_1 + \vec{s}_2, \vec{r})F_\alpha(\vec{t}_1 + \vec{t}_2 + \vec{t}_3, \vec{s}_1)F_\alpha(\vec{t}_4 + \vec{t}_5 + \vec{t}_6, \vec{s}_2)$$

$$F_\alpha(\vec{n}_1, \vec{t}_1)F_\alpha(\vec{n}_2, \vec{t}_2)F_\alpha(\vec{n}_3, \vec{t}_3)F_\alpha(\vec{n}_4, \vec{t}_4)F_\alpha(\vec{n}_5, \vec{t}_5)F_\alpha(\vec{n}_6, \vec{t}_6)$$

where

- $\vec{n}_1$, $\vec{n}_2$,... are the data at the leaf nodes, $\vec{t}_1$, $\vec{t}_2$,... their auxiliary counts
- $\vec{s}_1$ are auxiliary counts constrained by $\vec{t}_1 + \vec{t}_2 + \vec{t}_3$,
- $\vec{s}_2$ are auxiliary counts constrained by $\vec{t}_4 + \vec{t}_5 + \vec{t}_6$,
- $\vec{r}$ are auxiliary counts constrained by $\vec{s}_1 + \vec{s}_2$,

The Worked N-gram Style Model, cont.

Note the probabilities are then estimated from the auxiliary counts during MCMC. This is the standard recursive CRP formula.

$$\begin{aligned}
\widehat{\vec{\mu}} &= \frac{\vec{s}_1 + \vec{s}_2}{S_1 + S_2 + \alpha} + \frac{\alpha}{S_1 + S_2 + \alpha}\left(\frac{\vec{r}}{R + \alpha} + \frac{R}{R + \alpha}\frac{1}{L}\right) \\
\widehat{\vec{\theta}_1} &= \frac{\vec{t}_1 + \vec{t}_2 + \vec{t}_3}{T_1 + T_2 + T_3 + \alpha} + \frac{\alpha}{T_1 + T_2 + T_3 + \alpha}\widehat{\vec{\mu}} \\
\widehat{\vec{p}_1} &= \frac{\vec{n}_1}{N_1 + \alpha} + \frac{\alpha}{N_1 + \alpha}\widehat{\vec{\theta}_1}
\end{aligned}$$

Note in practice:

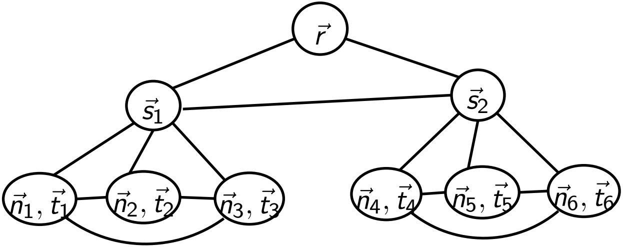
- the α is varied at every level of the tree and sampled as well,
- the PYP is used instead because words are often Zipfian

The Worked N-gram Style Model, cont.

What have we achieved:

- Bottom level probabilities $(\vec{p}_1, \vec{p}_2, ...)$ marginalised away.
- Each non-leaf probability vector $(\vec{\mu}, \vec{\theta}_1, ...)$ replaced by corresponding constrained auxiliary count vector $(\vec{r}, \vec{s}_1, ...)$ as psuedo-data.
- The auxiliary counts correspond to how much of the counts get inherited up the hierarchy.
- This allows a collapsed sampler in a discrete (versus continuous) space.

MCMC Problem Specification for N-grams



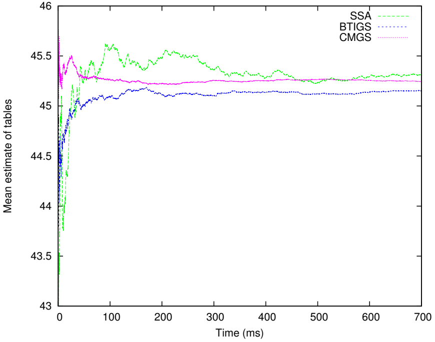
Build a Gibbs/MCMC sampler for:

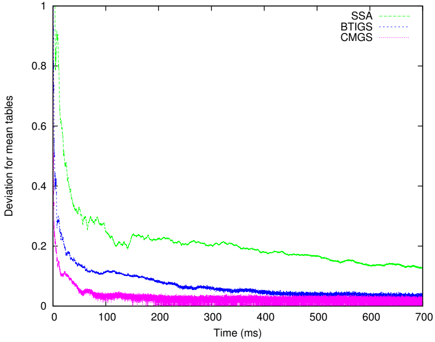
$$(\vec{\mu}) \quad \cdots \quad \frac{\alpha^R}{(\alpha)_{S_1+S_2}} \prod_{k=1}^{K} \left(\mathcal{S}^{s_{1,k}+s_{2,k}}_{r_k,0} \frac{1}{K^{r_k}} \right)$$

$$(\vec{\theta}_1, \vec{\theta}_2) \quad \cdots \quad \frac{\alpha^{S_1}}{(\alpha)_{T_1+T_2+T_3}} \prod_{k=1}^{K} \mathcal{S}^{t_{1,k}+t_{2,k}+t_{3,k}}_{s_{1,k},0} \frac{\alpha^{S_2}}{(\alpha)_{T_4+T_5+T_6}} \prod_{k=1}^{K} \mathcal{S}^{t_{4,k}+t_{5,k}+t_{6,k}}_{s_{2,k},0}$$

$$(\forall_k \vec{p}_k) \quad \cdots \quad \prod_{l=1}^{6} \left(\prod_{k=1}^{K} \mathcal{S}^{n_{l,k}}_{t_{l,k},0} \right)$$

CRP Samplers versus MultPYP, cont.

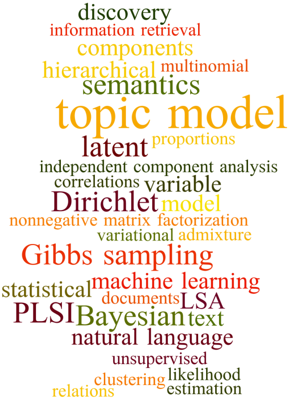




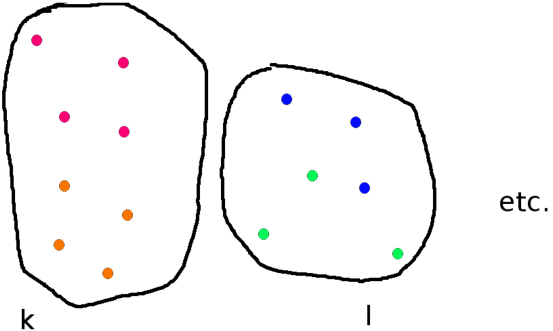
Legend: SSA = "standard CRP sampler of Teh *et al.*"
CMGS = "Gibbs sampler using MultPYP posterior"

Mean estimates of the total number of tables T for one of the 20 Gibbs runs (left) and the standard deviation of the 20 mean estimates (right) with $d = 0$, $\alpha = 10$, $K = 50$ and $N = 500$.

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

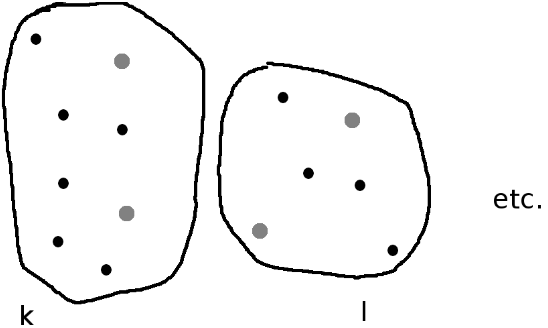
Species with Subspecies



Within species there are separate *sub-species*, pink and orange for type k, blue and green for type l.

Chinese restaurant samplers work in this space, keeping track of all counts for sub-species.

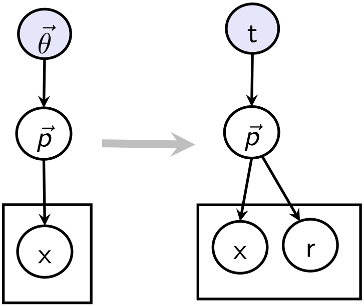
Species with New Species



Within species there are separate *sub-species*, but we only know which data is the first of a new sub-species.

Block table indicator samplers work in this space, where each datum has a Boolean indicator.

Categorical Data plus Table Indicators



LHS = *categorical* form with sample of discrete values $x_1, \ldots, x_N$ drawn from categorical distribution $\vec{p}$ which in turn has mean $\vec{\theta}$

RHS = *species sampling* form where data is now pairs $(x_1, r_1) \ldots, (x_N, r_N)$ were r_n is a Boolean indicator saying "is new subspecies"

$r_n = 1$ then the sample x_n was drawn from the parent node with probability θ_{x_n}, otherwise is existing subspecies

Table Indicators

Definition of table indicator

Instead of considering the multinomial-Pitman-Yor with counts $(\vec{n}, \vec{t})$, work with sequential data with individual values $(x_1, r_1), (x_2, r_2), \ldots, (x_N, r_N)$. The table indicator r_n indicates that the data contributes one count up to the parent probability.

So the data is treated sequentially, and taking statistics of $\vec{x}$ and $\vec{r}$ yields:

$$
\begin{aligned}
n_k &:= \text{counts of } k\text{'s in } \vec{x}, \\
&= \textstyle\sum_{n=1}^{N} 1_{x_n=k}, \\
t_k &:= \text{counts of } k\text{'s in } \vec{x} \text{ co-occuring with an indicator}, \\
&= \textstyle\sum_{n=1}^{N} 1_{x_n=k} 1_{r_n}.
\end{aligned}
$$

The Categorical-Pitman-Yor

Definition of Categorical-Pitman-Yor

Given a concentration parameter α, a discount parameter d, a probability vector $\vec{\theta}$ of dimension L, and a count N, the categorical-Pitman-Yor distribution creates a sequence of discrete class assignments and indicators $(x_1, r_1), ...(x_N, r_N)$. Now $(\vec{x}, \vec{r}) \sim \text{CatPYP}\left(d, \alpha, \vec{\theta}, N\right)$ denotes

$$p\left(\vec{x}, \vec{r} \,\middle|\, N, \text{CatPYP}, d, \alpha, \vec{\theta}\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{l=1}^{L} \mathcal{S}^{n_l}_{t_l,d} \theta_l^{t_l} \binom{n_l}{t_l}^{-1}$$

where the counts are derived, $t_l = \sum_{n=1}^{N} 1_{x_n=l} 1_{r_l}$, $n_l = \sum_{n=1}^{N} 1_{x_n=l}$, $T = \sum_{l=1}^{L} t_l$.

The Categorical- versus multinomial-Pitman-Yor

Multinomial-Pitman-Yor: working off counts $\vec{n}, \vec{t}$,

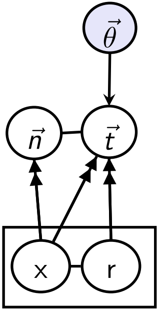
$$p\left(\vec{n}, \vec{t} \,\middle|\, N, \text{MultPYP}, d, \alpha, \vec{\theta}\right) = \binom{N}{\vec{n}} \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} \theta_k^{t_k}$$

Categorical-Pitman-Yor: working off sequential data $\vec{x}, \vec{r}$, the counts $\vec{n}, \vec{t}$ are now derived,

$$p\left(\vec{x}, \vec{r} \,\middle|\, N, \text{CatPYP}, d, \alpha, \vec{\theta}\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} \theta_k^{t_k} \binom{n_k}{t_k}^{-1}$$

- remove the $\binom{N}{\vec{n}}$ term because sequential order now matters
- divide by $\binom{n_k}{t_k}$ because this is the number of ways of distributing the t_k indicators that are on amongst n_k places

Pitman-Yor-Categorical Marginalisation

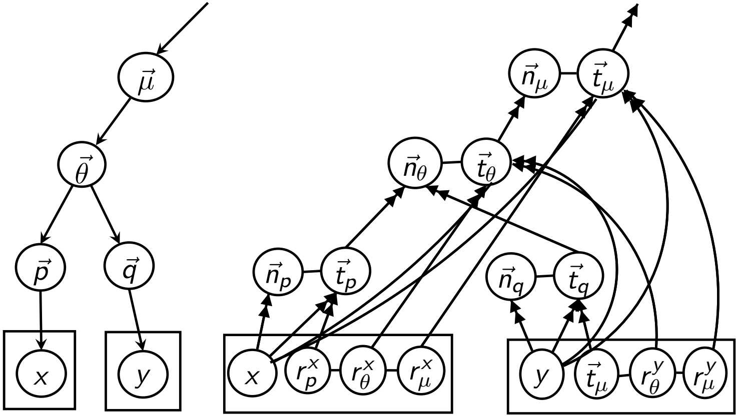


$\vec{n} =$ *vector of counts* of different species (how much data of each species); computed from the data $\vec{x}$

$\vec{t} =$ count vector giving how many different subspecies; computed from the paired data $\vec{x}, \vec{r}$; called *number of tables*

$$p\left(\vec{x}, \vec{r} \,\middle|\, d, \alpha, \mathrm{PYP}, \vec{\theta}\right) = \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \theta_k^{t_k} \mathcal{S}^{n_k}_{t_k,d} \binom{n_k}{t_k}^{-1}$$

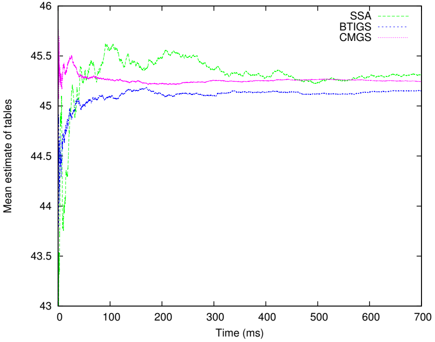
Hierarchical Marginalisation

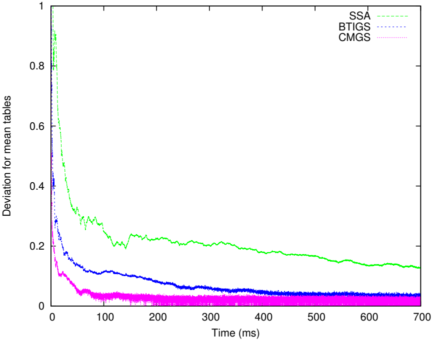


left is the original probability vector hierarchy, **right** is the result of marginalising out probability vectors then

- indicators are attached to their originating data as a set
- all $\vec{n}$ and $\vec{t}$ counts up the hierarchy are computed from these

CRP Samplers versus MultPYP, cont.





Legend: SSA = "standard CRP sampler of Teh *et al.*"
CMGS = "Gibbs sampler using MultPYP posterior"

Mean estimates of the total number of tables T for one of the 20 Gibbs runs (left) and the standard deviation of the 20 mean estimates (right) with $d = 0$, $\alpha = 10$, $K = 50$ and $N = 500$.

Better Sampling Methods for HDP and HPYP

Sampling for hierarchical Dirichlet Processes and Pitman-Yor Processes:

The Old: hierarchical Chinese Restaurant Processes (CRP) from Teh *et al.* 2006.

The New: block table indicator sampling from Chen, Du and Buntine 2011.

- requires no dynamic memory
- more rapid mixing so leads to better models
- more easily applied to more complex models
- demonstrated extensively on different problems!

See `http://topicmodels.org` , "A tutorial on non-parametric methods"

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Conjugate Discrete Families

Conjugate Family	$p(x\|\lambda)$	$p(\lambda) \propto$
Bernoulli-Beta	$\lambda^x(1-\lambda)^{1-x}$	$\lambda^{\alpha-1}(1-\lambda)^{\beta-1}\delta_{0<\lambda<1}$
Poisson-Gamma	$\frac{1}{x!}\lambda^x e^{-\lambda}$	$\lambda^{\alpha-1}e^{-\beta\lambda}$
negtve-Binomial-Gamma	$\frac{1}{x!}(\lambda)_x\,\rho^x(1-\rho)^\lambda$	$\lambda^{\alpha-1}e^{-\beta\lambda}$

parameters $\alpha, \beta > 0$
$(\lambda)_x$ is rising factorial $\lambda(\lambda+1)...(\lambda+x-1)$

- multinomial-Dirichlet used in LDA
- Poisson-Gamma in some versions of NMF
- Bernoulli-Beta is the basis of IBP
- negative-Binomial-Gamma is not quite a conjugate family; the negative-Binomial is a "robust" variant of a Poisson

Infinite Vectors

- Let $\omega_k \in \Omega$ for $k = 1, ..., \infty$ be index points for an infinite vector.
- Have infinite parameter vector $\vec{\lambda} = \sum_{k=1}^{\infty} \lambda_k \delta_{\omega_k}$ for $\lambda_k \in (0, \infty)$,
- Generate I discrete feature vectors $\vec{x}_i = \sum_{k=1}^{\infty} x_{i,k} \delta_{\omega_k}$ pointwise using discrete distribution $p(x_{i,k}|\lambda_k)$.
- Require only finite number of $x_{i,k} \neq 0$ for given i.
- This means we need $\sum_{k=1}^{\infty} \lambda_k < \infty$,
  - assuming lower λ_k makes $x_{i,k}$ more likely to be zero.
- Can arbitrarily rearrange dimensions k since all objects have term $\sum_{k=1}^{\infty}(\cdot)$.
- Don't know which dimensions k non-zero so rearrange as needed.

See 2014 ArXiv paper by Lancelot James (`http://arxiv.org/abs/1411.2936`).

Generating Infinite Vectors

- Almost all λ_k must be infinitesimally small, so the prior for λ is not proper. Formally modelled using Poisson processes:

  Have Poisson process with points λ, ω on domain $(0, \infty) \times \Omega$ with rate $p(\lambda|\omega) d\lambda G(d\omega)$.

- For infinite number of points want (cannot normalise)

$$\int_0^\infty \int_\Omega p(\lambda|\omega) \mathrm{d}\,\lambda G(\mathrm{d}\,\omega) = \infty$$

- To expect $\sum_{k=1}^\infty \lambda_k < \infty$, want

$$\int_0^\infty \int_\Omega \lambda p(\lambda|\omega) \mathrm{d}\,\lambda G(\mathrm{d}\,\omega) < \infty$$

- Parameters to the "distribution" (Poisson process rate) for λ would control the expected number of non-zero $x_{i,k}$.

Bernoulli-Beta Process (Indian Buffet Process)

- infinite Boolean vectors $\vec{x}_i$ with a finite number of 1's;
- each parameter λ_k is an independent probability,

$$p(x_{i,k}|\lambda_k) \;=\; \lambda_k^{x_{i,k}}(1-\lambda_k)^{1-x_{i,k}}$$

- to have finite 1's, require $\sum_k \lambda_k < \infty$
- improper prior (Poisson process rate) is the 3-parameter Beta process

$$p(\lambda|\alpha,\beta,\theta) \;=\; \theta\lambda^{-\alpha-1}(1-\lambda)^{\alpha+\beta-1}$$

(some versions add additional constants with θ)

- is in improper Beta because seeing "1" makes it proper:

$$\int_{\lambda=0}^{1} p(x=1|\lambda)p(\lambda)\mathrm{d}\,\lambda \;=\; \theta\,\mathrm{Beta}(1-\alpha,\alpha+\beta)$$

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Conjugate Discrete Processes

Each conjugate family has a corresponding non-parametric version:

- Uses the improper versions of the prior $p(\lambda|\omega)$
  e.g. for Gamma, Beta, Dirichlet
- Want to generate a countably infinite number of λ but have almost all infinitesimally small.
- Theory done with Poisson processes, see 2014 ArXiv paper by Lancelot James (`http://arxiv.org/abs/1411.2936`).
- Presention here uses the more informal language of “improper priors,” but the correct theory is Poisson processes.

Conjugate Discrete Processes, cont.

Non-parametric versions of models for discrete feature vectors:

Process Name	$p(x\|\lambda)$	$p(\lambda)$
Poisson-Gamma	$\frac{1}{x!}\lambda^x e^{-\lambda}$	$\theta\lambda^{-\alpha-1}e^{-\beta\lambda}$
Bernoulli-Beta	$\lambda^x(1-\lambda)^{1-x}$	$\theta\lambda^{-\alpha-1}(1-\lambda)^{\alpha+\beta-1}\delta_{0<\lambda<1}$
negtve-Binomial-Gamma	$\frac{1}{x!}(\lambda)_x\rho^x(1-\rho)^\lambda$	$\theta\lambda^{-\alpha-1}e^{-\beta\lambda}$

$\beta, \theta > 0$
$0 \leq \alpha < 1$

- In common they make the power of λ lie in $(-2, -1]$ to achieve the "improper prior" effect.
- Term θ is just a general proportion to uniformally increase number of λ_k's in any region.
- Whereas α and β control the relative size of the λ_k's.

Conjugate non-Parametric Discrete Families, cont.

- Given λ, probability of I samples with at least one non-zero entry is

$$\left(1 - p(x_{i,k} = 0|\lambda)^I\right) .$$

- By Poisson process theory, expectation of this (in general case)

$$\Psi_I = \int_\Omega \int_0^\infty \left(1 - p(x_{i,k} = 0|\lambda)^I\right) \rho(\lambda|\omega) \mathrm{d}\,\lambda G_0(\mathrm{d}\,\omega_k)$$

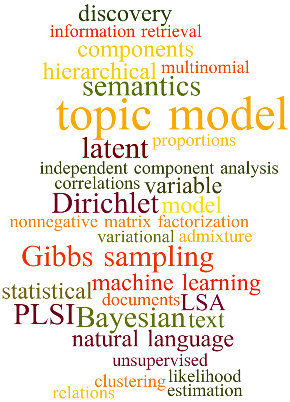
- Call Ψ_I the Poisson non-zero rate, a function of I and the underlying distributions.
- With I vectors, number of non-zero dimensions K is Poisson with rate Ψ_I, having probability

$$\frac{1}{K!} e^{-\Psi_I} \Psi_I^K .$$

Posterior Marginal

- With I vectors, number of non-zero dimensions K is Poisson with rate Ψ_I, having probability
$$\frac{1}{K!} e^{-\Psi_I} \Psi_I^K \ .$$
- Take particular dimension ordering (remove $\frac{1}{K!}$) and replace "not all zero" by actual data, $x_{i,1}, ..., x_{i,K}$ to get:
$$e^{-\Psi_I} \prod_{k=1}^{K} p(x_{1,k}, ..., x_{i,k}, \omega_w) \ .$$
- Expand using model to get posterior marginal:
$$p(\vec{x}_1, ..., \vec{x}_I, \vec{\omega}) = e^{-\Psi_I} \prod_{k=1}^{K} \left(\int_0^\infty \left(\prod_{i=1}^{I} p(x_{i,k}|\lambda) \right) \rho(\lambda|\omega) \mathrm{d}\,\lambda \right) G_0(\mathrm{d}\,\omega_k)$$

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Bernoulli-Beta Process (Indian Buffet Process)

- The Poisson non-zero rate
  trick: use $1 - y^I = (1-y)\sum_{i=0}^{I} y^i$

$$\Psi_I \;=\; \theta\Gamma(1-\alpha)\sum_{i=0}^{I}\frac{\Gamma(\beta+\alpha+i)}{\Gamma(\beta+1+i)}\;.$$

- The marginal for the k-th dimension

$$\int_0^\infty\left(\prod_{i=1}^{I}p(x_{i,k}|\lambda)\right)\rho(\lambda|\omega)\mathrm{d}\,\lambda \;=\; \theta\,\mathrm{Beta}(c_k-\alpha, I-c_k+\alpha+\beta)$$

  where c_k is times dimension k is "on," so $c_k = \sum_{i=1}^{I} x_{i,k}$.
- Gibbs sampling $x_{i,k}$ is thus simple.
- Sampling parameters: posterior of θ is Poisson; posterior for β is log-concave so sampling "easier".

Poisson-Gamma Process

- The Poisson non-zero rate
  trick: use the Laplace exponent from Poisson process theory

$$\Psi_I = \theta \frac{\Gamma(1-\alpha)}{\alpha}\left((I+\beta)^\alpha - \beta^\alpha\right) .$$

- The marginal for the k-th dimension

$$\int_0^\infty \left(\prod_{i=1}^I p(x_{i,k}|\lambda)\right)\rho(\lambda|\omega)\mathrm{d}\,\lambda = \theta\left(\prod_{i=1}^I \frac{1}{x_{i,k}!}\right)\frac{\Gamma(x_{\cdot,k}-\alpha)}{(I+\beta)^{x_{\cdot,k}-\alpha}}$$

  where $x_{\cdot,k} = \sum_{i=1}^I x_{i,k}$.
- Gibbs sampling the $x_{i,k}$ is thus simple.
- Sampling parameters: posterior of θ is Poisson; posterior of β is unimodal (and no other turning points) with simple closed form for MAP.

Negative-Binomial-Gamma Process

- Series of papers for this case by Mingyuan Zhou and colleages.
- The Poisson non-zero rate
  trick: use the Laplace exponent from Poisson process theory

$$\Psi_I = \theta \frac{\Gamma(1-\alpha)}{\alpha}\left(\left(I \log\left(\frac{1}{1-p}\right)+\beta\right)^{\alpha} - \beta^{\alpha}\right) .$$

- The marginal for the k-th dimension

$$\int_0^\infty \left(\prod_{i=1}^I p(x_{i,k}|\lambda,p)\right)\rho(\lambda|\omega)\mathrm{d}\,\lambda$$

$$= \quad p^{x_{\cdot,k}}\left(\prod_{i=1}^I \frac{1}{x_{i,k}!}\right)\int_0^\infty (1-p)^{I\lambda}\left(\prod_{i=1}^I (\lambda)_{x_{i,k}}\right)\rho(\lambda)\mathrm{d}\,\lambda$$

- Gibbs sampling the $x_{i,k}$ is more challenging.
  - keep λ as a latent variable (posterior is log concave);
  - use approximation $(\lambda)_x \approx \lambda^{t^*} S^x_{t^*,0}$ where $t^* = \mathrm{argmax}_{t\in[1,x)}\lambda^t S^x_{t,0}$.

Simple, Fast Hierarchical IBP

James' more general theory allows more creativity in construction.

Bernoulli-Beta-Beta process

- model is a hierarchy of Bernoulli-Beta processes
- infinite feature vector $\vec{\lambda}$ is a Beta Process as before;
- these varied with point-wise Beta distributions to create a set of parent nodes $\vec{\psi}_j$, so $\psi_{j,k} \sim \text{Beta}(\alpha\lambda_{j,k}, \alpha(1-\lambda_{j,k}))$
- discrete features ordered in a hierarchy below nodes j so $x_{i,k} \sim \text{Bernoulli}(\psi_{j,k})$ for j the parent of node i.
- Use hierarchical Dirichlet process techniques to implement efficiently.

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Component Models, Generally

image →

Prince, Queen, Elizabeth, title, son, ...	school, student, college, education, year, ...
John, David, Michael, Scott, Paul, ...	and, or, to , from, with, in, out, ...

text →

13 1995 accompany and(2) andrew at boys(2) charles close college day despite diana dr eton first for gayley harry here housemaster looking old on on school separation sept stayed the their(2) they to william(2) with year

Approximate faces/bag-of-words (RHS) with a linear combination of components (LHS).

Matrix Approximation View

$$\boldsymbol{W} \simeq \boldsymbol{L} * \boldsymbol{\Theta}^T$$

$$\text{I documents}\left\{ \overbrace{\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,J} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,2} & \cdots & w_{I,J} \end{pmatrix}}^{J \text{ words}} \right. \simeq \overbrace{\begin{pmatrix} l_{1,1} & \cdots & l_{1,K} \\ l_{2,1} & \cdots & l_{2,K} \\ \vdots & \ddots & \vdots \\ l_{I,1} & \cdots & l_{I,K} \end{pmatrix}}^{K \text{ components}} * \left. \overbrace{\begin{pmatrix} \theta_{1,1} & \theta_{2,1} & \cdots & \theta_{J,1} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1,K} & \theta_{2,K} & \cdots & \theta_{J,K} \end{pmatrix}}^{J \text{ words}} \right\} \text{K components}$$

Different variants:

Data $\boldsymbol{W}$	Components $\boldsymbol{L}$	Error	Models
real valued	unconstrained	least squares	PCA and LSA
non-negative	non-negative	least squares	learning codebooks, NMF
non-neg integer	non-negative	cross-entropy	topic modelling, NMF
real valued	independent	small	ICA

Why Topic Models?

- *Topic Models* discover hidden themes in text data to aid understanding.
  - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003).
- Recent research develops higher performance topic models.

Why Topic Models?

- *Topic Models* discover hidden themes in text data to aid understanding.
    - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003).
- Recent research develops higher performance topic models.
- But why should you care?
- Moreover, why should I care?

Topic Models: Potential for Semantics

Following sets of topic words created from the New York Times 1985-2005 news collection using `hca` (see Buntine and Mishra, KDD 2014):

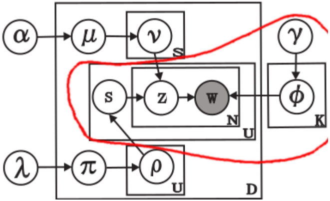
- career,born,grew,degree,earned,graduated,became,studied,graduate
- mother,daughter,son,husband,family,father,parents,married,sister
- artillery,shells,tanks,mortars,gunships,rockets,firing,tank
- clues,investigation,forensic,inquiry,leads,motive,investigator,mystery
- freedom,tyranny,courage,america,deserve,prevail,evil,bless,enemies
- viewers,cbs,abc,cable,broadcasting,channel,nbc,broadcast,fox,cnn
- anthrax,spores,mail,postal,envelope,powder,letters,daschle,mailed

Topic models yield high-fidelity semantic associations!

Topic Models: Just an Intermediate Goal

Topic models are the leading edge of a new wave of deep latent semantic models applied to real NLP tasks:

e.g., document segmentation, word sense disambiguation, facet discovery for sentiment analysis, unsupervised POS discovery, social networks, ...



in the middle of this segmentation model is a topic model

i.e., we don't care about topic models *per se*!

ASIDE: Aspects, Ratings and Sentiments

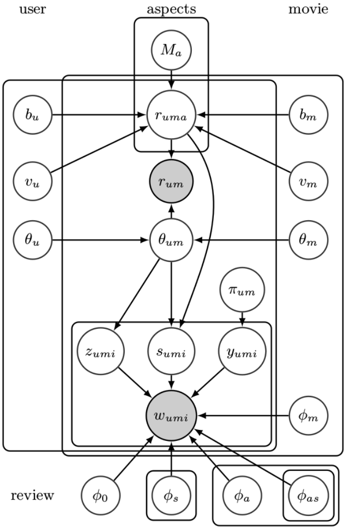


Figure 1: Factorized rating and review model.

"Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS)," Diao, Qiu, Wu, Smola, Jiang and Wang, KDD 2014.

State of the art sentiment model.

Typical methods currently lack probability vector hierarchies.

Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test” (not his exact words .... but the same idea)

Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test” (not his exact words .... but the same idea)

Perplexity:

- measure of test set likelihood;
- equal to effective size of vocabulary;
- we use “document completion,” see Wallach, Murray, Salakhutdinov, and Mimno, 2009;
- however it is not a bonafide evaluation task

Evaluation

David Lewis (Aug 2014) “topic models are like a Rorschach inkblot test” (not his exact words .... but the same idea)

Perplexity:
- measure of test set likelihood;
- equal to effective size of vocabulary;
- we use “document completion,” see Wallach, Murray, Salakhutdinov, and Mimno, 2009;
- however it is not a bonafide evaluation task

PMI:
- measure of topic coherence: *“average pointwise mutual information between all pairs of top 10 words in the topic”*
- see Newman, Lau, Grieser, and Baldwin, 2010; Lau, Newman and Baldwin, 2014
- but at least it corresponds to a semi-realistic evaluation task

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Previous Work

- "Hierarchical Dirichlet Processes," Teh, Jordan, Beal, Blei 2006.
- "Rethinking LDA: Why priors matter," Wallach, Mimno, McCallum, 2009.
- "Accounting for burstiness in topic models," Doyle and Elkan 2009.
- "Topic models with power-law using Pitman-Yor process," Sato and Nakagawa 2010
- Sampling table configurations for the hierarchical Poisson-Dirichlet process," Chen, Du and Buntine 2011.
- "Practical collapsed variational Bayes inference for hierarchical Dirichlet process," Sato, Kurihara, and Nakagawa 2012.
- "Truly nonparametric online variational inference for hierarchical Dirichlet processes," Bryant and Sudderth 2012.
- "Stochastic Variational Inference," Hoffman, Blei, Wang and Paisley 2013.
- "Latent IBP compound Dirichlet Allocation," Archambeau, Lakshminarayanan, Bouchard 2014.

Text and Burstiness

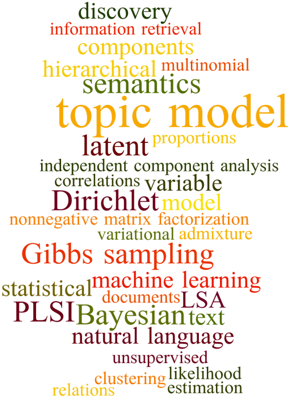
Original news article:	Women may only account for 11% of all Lok-Sabha MPs but they fared better when it came to representation in the Cabinet. Six women were sworn in as senior ministers on Monday, accounting for 25% of the Cabinet. They include Swaraj, Gandhi, Najma, Badal, Uma and Smriti.
Bag of words:	11% 25% Badal Cabinet(2) Gandhi Lok-Sabha MPs Monday Najma Six Smriti Swaraj They Uma Women account accounting all and as better but came fared for(2) in(2) include it may ministers of on only representation senior sworn the(2) they to were when women

NB. "Cabinet" appears twice! It is **bursty**
(see Doyle and Elkan, 2009)

Aside: Burstiness and Information Retrieval

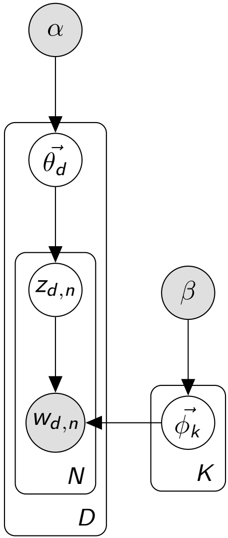
- burstiness and eliteness are concepts in information retrieval used to develop BM25 (*i.e.* dominant TF-IDF version)
- the two-Poisson model and the Pitman-Yor model can be used to justify theory (Sunehag, 2007; Puurula, 2013)
- relationships not yet fully developed

Outline



1. Background
2. Non-parametric Bayesian Methods
3. Discrete Feature Vectors
4. High Performance Topic Models (with Swapnil Mishra)
   - Topic Models
   - Background
   - Evolution of Models
   - Our Non-parametric Topic Model
   - Experimental Comparisons
5. Twitter Opinion Topic Model (with Kar Wai

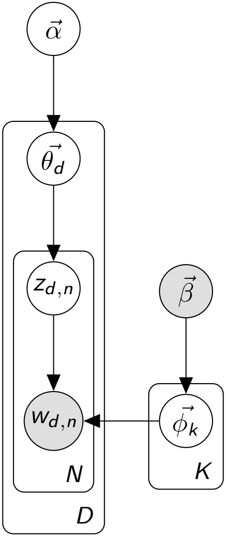
Evolution of Models



LDA- Scalar

original LDA

Evolution of Models



LDA- Vector

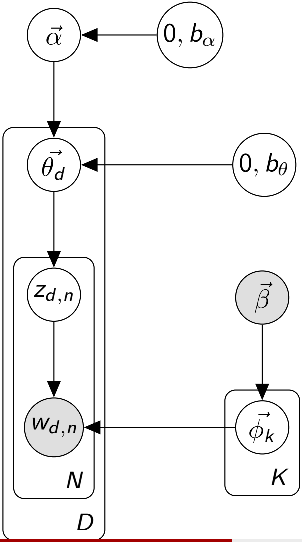
adds asymmetric Dirichlet prior like Wallach et al.;

is also truncated HDP-LDA;

implemented by Mallet since 2008 as assymetric-symmetric LDA

no one knew!

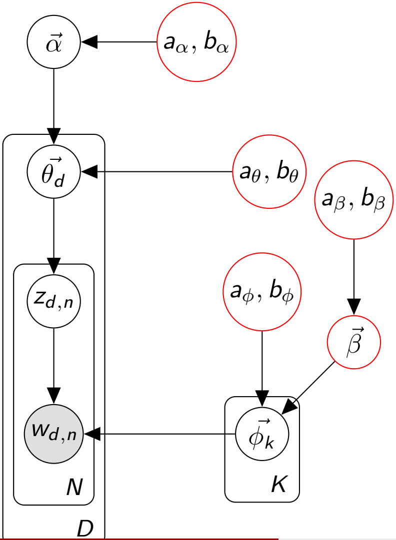
Evolution of Models



HDP-LDA

adds proper modelling of topic prior like Teh et al.

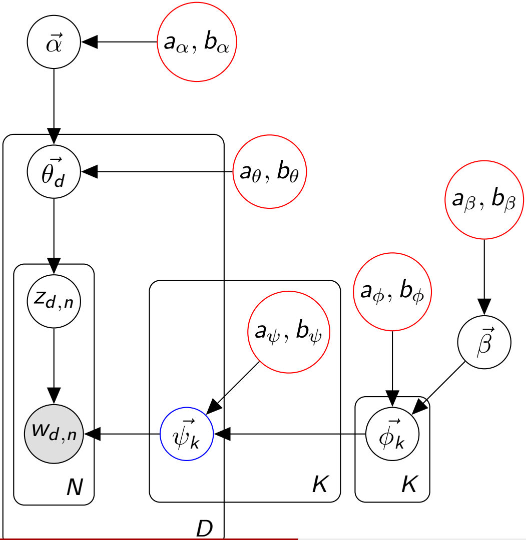
Evolution of Models



NP-LDA

adds power law on word distributions like Sato et al. and estimation of background word distribution

Evolution of Models



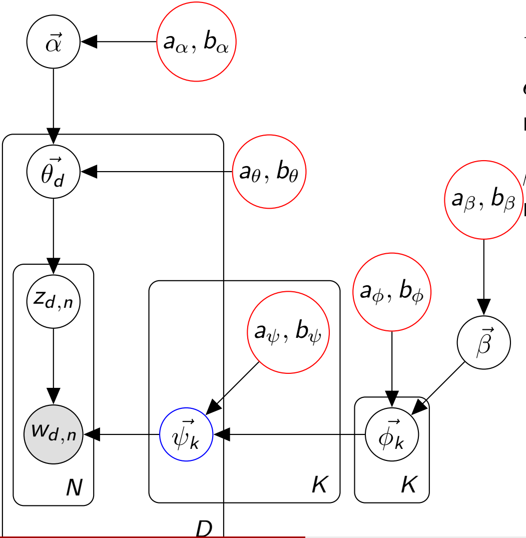
NP-LDA with Burstiness

add's burstiness like Doyle and Elkan

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Our Non-parametric Topic Model



$\{\vec{\theta}_d\}$ = document⊗topic matrix

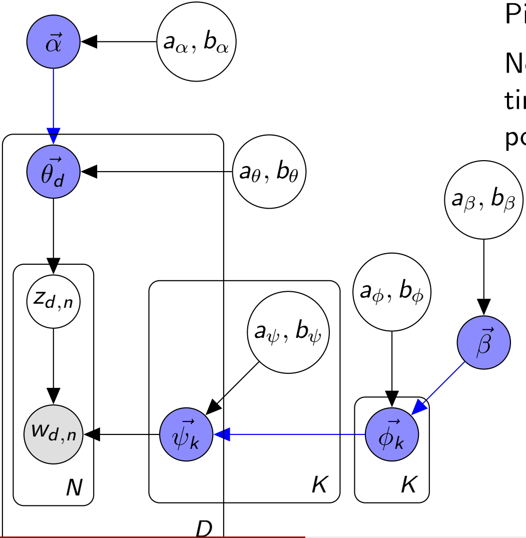
$\{\vec{\phi}_k\}$ = topic⊗word matrix

$\vec{\alpha}$ = prior for document⊗topic matrix

$\vec{\beta}$ = prior for topic⊗word matrix

- Full fitting of priors, and their hyperparameters.
- Topic⊗word vectors $\vec{\phi}_k$ specialised to the document to yield $\vec{\psi}_k$.

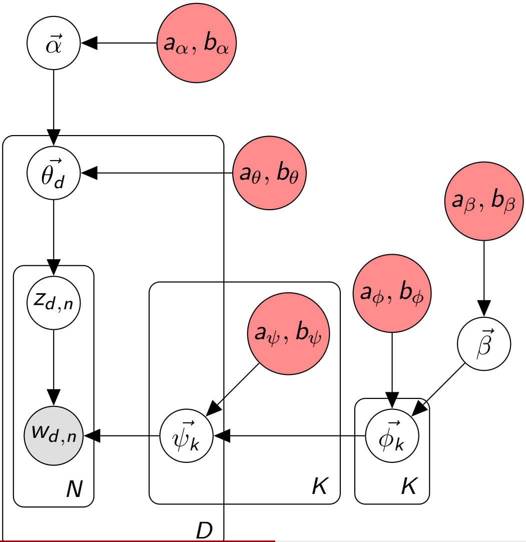
Our Non-parametric Topic Model, cont.



The blue nodes+arcs are Pitman-Yor process hierarchies.

Note in $\{\vec{\psi}_k\}$ there are hundreds times more parameters than data points!

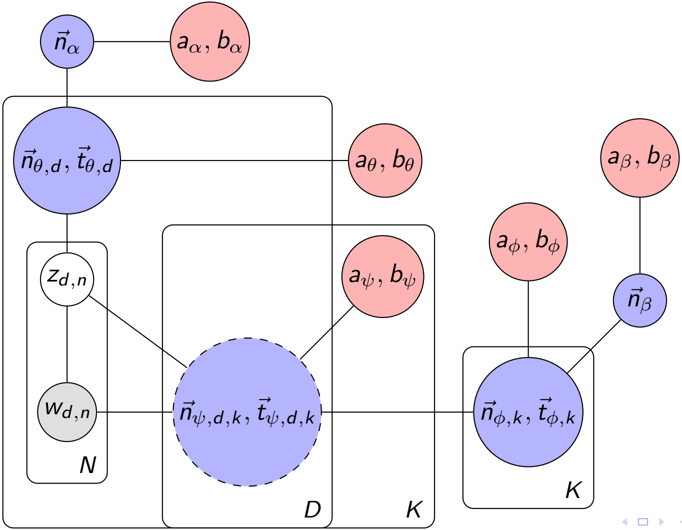
Our Non-parametric Topic Model, cont.



The red nodes are hyper-parameters fit with Adaptive-Rejection sampling or slice sampling.

Use DP on document side ($a_\alpha = 0, a_\theta = 0$) as fitting usually wants this anyway.

Our Non-parametric Topic Model, cont.



Auxiliary latent variables (the $\vec{t}$) propagate part of the counts (their $\vec{n}$) up to the parent.

We keep/recompute sufficient statistics for matrices.

e.g. the $\vec{\psi}$ statistics $\vec{n}_{\psi,d,k}, \vec{t}_{\psi,d,k}$ are not stored but recomputed from booleans as needed.

Double the memory of regular LDA, and only static memory.

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Estimating parameters: whenever parameters cannot be reasonable set, we learn them instead.

Design Notes

Hierarchical priors: whenever parts of the system seem similar, we give them a common prior and learn the similarity.

Estimating parameters: whenever parameters cannot be reasonable set, we learn them instead.

Burstiness: we developed a Gibbs sampler that acts as a front end to **any** LDA-style model with Gibbs:

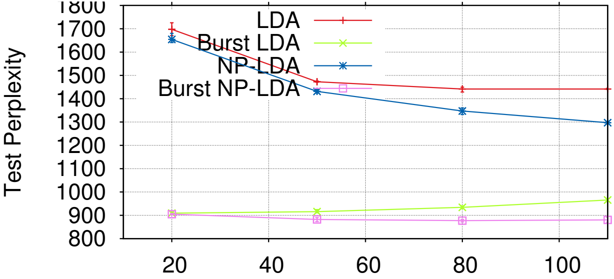
- implemented as a C function that calls the Gibbs sampler
- adds smallish memory (20%) and time (20%) overhead
- in all, NP-LDA with burstiness is double memory and time to regular LDA Gibbs sampling
- multi-core implementation good for upto 8 core

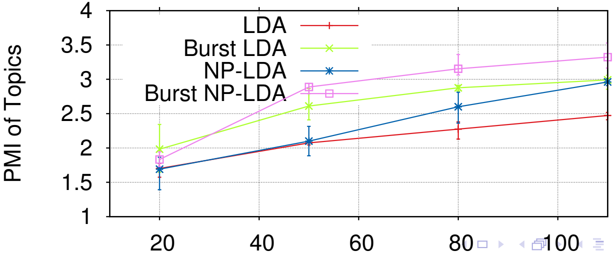
Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Performance on Reuters-21578 ModLewis Split

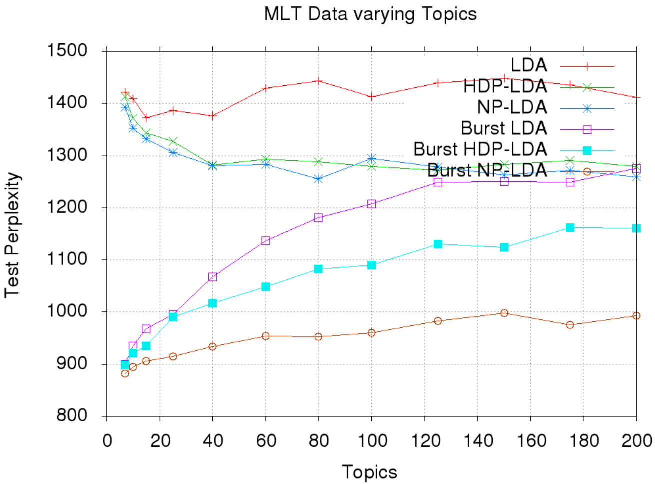
Training on 11314 news articles with vocabulary of 16994.



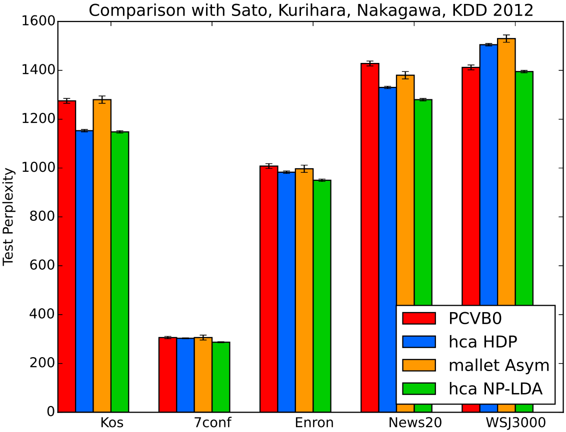


Perplexity performance on MLT Data for different Topics

2691 abstracts from the JMLR including 306 test documents with a vocabulary of 4662 words



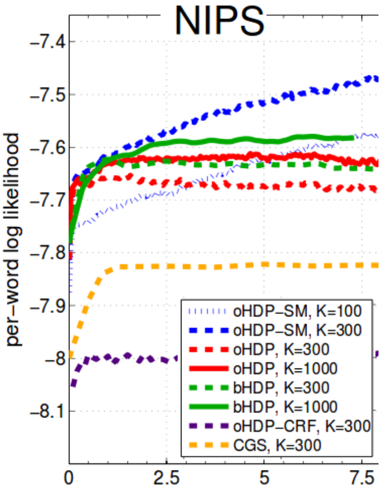
Comparison to PCVB0 and Mallet

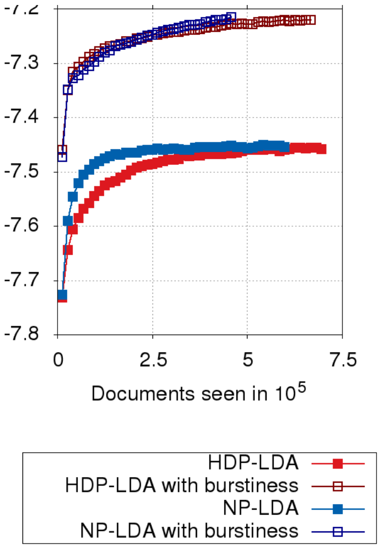


Protocol is train on 80% of all documents then using trained topic probs get predictive probabilities on remaining 20%, and replicate 5 times.

- Data contributed by Sato. Protocol by Sato *et al.*
- PCVB0 is by Sato, Kurihara, Nakagawa KDD 2012.
- Mallet (asymmetric-symmetric) is a truncated HDP implementation.

Comparison to Bryant+Sudderth (2012) on NIPS data





Comparison to FTM and LIDA

FTM and LIDA use IBP models to select words/topics within LDA.
Archambeau, Lakshminarayanan, and Bouchard, *Trans IEEE PAMI* 2014.

Data	KOS	NIPS
FTM (1-par)	7.262±0.007	6.901±0.005
FTM (3-par)	7.266±0.009	6.883±0.008
LIDA	7.257±0.010	6.795±0.007
HPD-LDA	7.253±0.003	6.792±0.002
time	3 min	22 min
NP-LDA	7.156±0.003	6.722±0.003

- KOS data contributed by Sato (D=3430, V=6906). NIPS data from UCI (D=1500, V=12419).
- Protocol same as with PCVB0 but a 50-50 split. Figures are log perplexity. Using 300 cycles.

- Better implementation of HDP-LDA now similar to LIDA.
- But LIDA still substantially better than LDA so we need to consider combining the technique with NP-LDA.

Conclusion on Topic Models

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this

Conclusion on Topic Models

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models

Conclusion on Topic Models

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art algorithms.
  - simple (4-8 cpu) multicore version available

Conclusion on Topic Models

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art algorithms.
  - simple (4-8 cpu) multicore version available
- Still need to explore IBP (as in LIDA) and split-merge techniques.

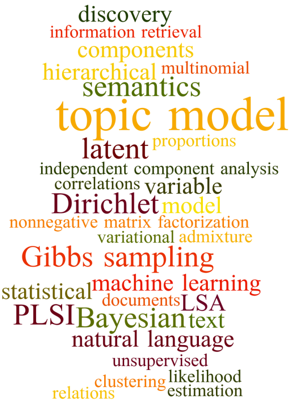
Conclusion on Topic Models

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art algorithms.
  - simple (4-8 cpu) multicore version available
- Still need to explore IBP (as in LIDA) and split-merge techniques.
- Grab our topic modelling code from

  https://github.com/wbuntine/topic-models
  http://mloss.org/software/view/527/

See KDD 2014 paper by Mishra and Buntine.

Outline

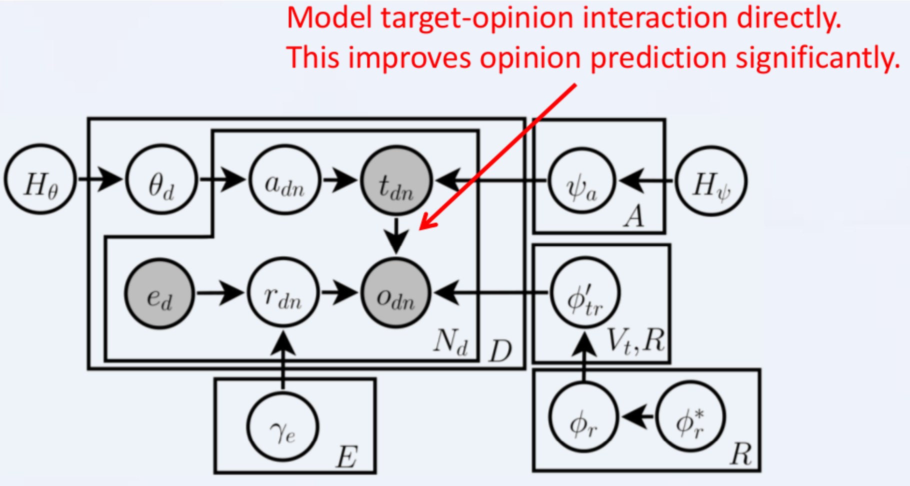

discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Aspect-based Opinion Aggregation

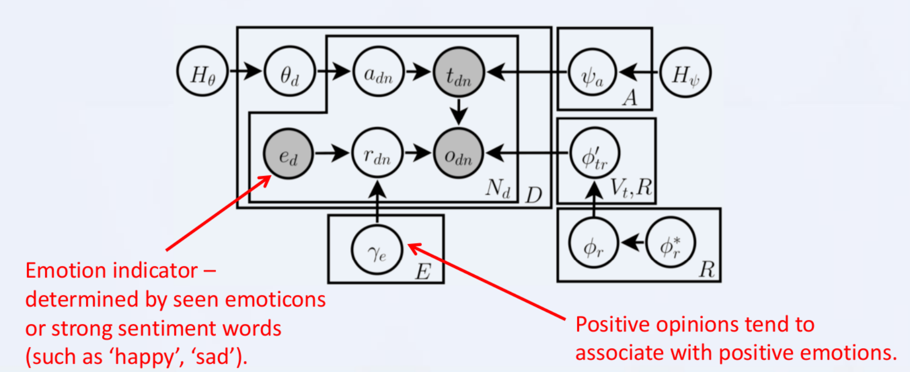
- Opinion Aggregation for reviews.
  - A process to collect reviews of products and services to analyze in aggregate.
- Aspect-based.
  - Groups reviews based on “aspects”.
  - Example:
    - Product types
      - Game consoles
      - Mobile phones
    - Product specs
      - Computer specs
      - Flight quality

Aspect	Examples
Game console	PS4, Xbox One, Wii U...
Mobile phone	iPhone, Samsung Note...
Computer spec	CPU, RAM, GPU...
Flight quality	Food, customer service...

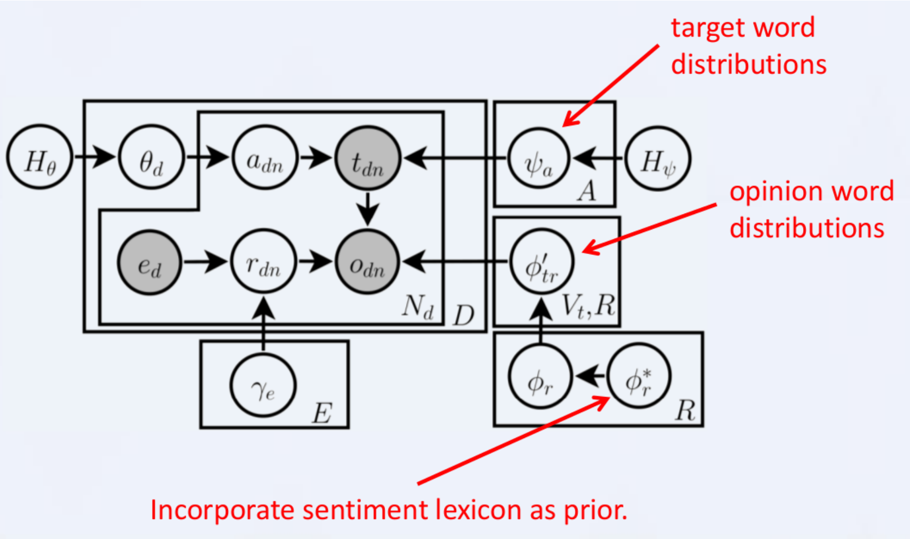
Explaining the Model



Explaining the Model, cont.



Explaining the Model, cont.



Explaining the Model, cont.

Target (t)	+/−	Opinions (o)
phone	−	**dead** damn stupid bad crazy
	+	**mobile smart** good great f***ing
battery life	−	terrible poor bad horrible **non-existence**
	+	good **long** great **7hr ultralong**
game	−	**addictive** stupid free **full addicting**
	+	great good awesome favorite **cat-and-mouse**
sausage	−	silly **argentinian cold** huge stupid
	+	**hot grilled** good **sweet** awesome

* Words in **bold** are more specific and can only describe certain targets.

Outline


discovery
information retrieval
components
hierarchical
multinomial
semantics
topic model
latent
proportions
independent component analysis
correlations
variable
Dirichlet
model
nonnegative matrix factorization
variational
admixture
Gibbs sampling
statistical
machine learning
documents
LSA
PLSI
Bayesian
text
natural language
unsupervised
clustering
likelihood
relations
estimation

Task, Roughly

Unsegmented Text

The EPA has issued a rule requiring greater use of ethanol. Made from corn, ethanol helps gasoline burn cleaner and lowers some kinds of emissions. Well, if you drive a car, you'll really be interested in our next report. New rules about gasoline mean cleaner air but higher prices at the pump. CNN's Deborah Potter explains. The government is opening the way for more of this *[corn]* to wind up, not in the trough or on the table, but here, at the gas pump. To clear the air in the nation's smoggiest cities, the EPA has ordered the use of cleaner-burning gasoline. Adding renewable fuels to gasoline promotes products that are grown on American farms by American farmers. It promotes environmentally friendly jobs. It reduces air pollution. It protects the public's health-

Segmented Text

The EPA has issued a rule requiring greater use of ethanol. Made from corn, ethanol helps gasoline burn cleaner and lowers some kinds of emissions.

Well, if you drive a car, you'll really be interested in our next report. New rules about gasoline mean cleaner air but higher prices at the pump. CNN's Deborah Potter explains.

The government is opening the way for more of this *[corn]* to wind up, not in the trough or on the table, but here, at the gas pump. To clear the air in the nation's smoggiest cities, the EPA has ordered the use of cleaner-burning gasoline.

Adding renewable fuels to gasoline promotes products that are grown on American farms by American farmers. It promotes environmentally friendly jobs. It reduces air pollution. It protects the public's health-

Task, Roughly

Passage: contiguous text with no boundary, *e.g.*, a sentence

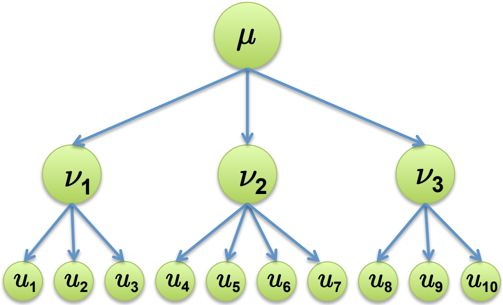
Segment: consecutive text passages that are semantically related.

Document: a sequence of topically coherent text segments.

Document Segmentation Task: (roughly) given a document as a monolithic block of text, where should we put the segment boundaries.

Motivation–Structured Topic Modelling

Structured topic models (STM) by Du et al., (2010): hierarchical topic models with non-parametric Bayesian methods.

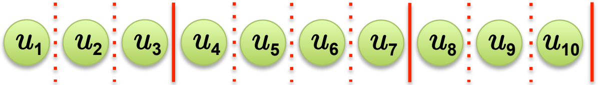


Has a hierarchy of topic probability vectors corresponding to document structure.

Bayesian Segmentation

Bayesian word segmentation models (Goldwater et al., 2009)

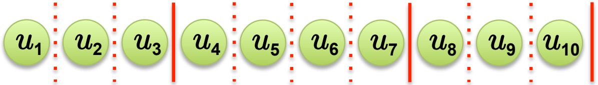
- Learn to place boundaries after phonemes in an utterance.
- A pointwise boundary sampling algorithm: compute the probability of placing a word boundary after each phoneme.



Bayesian Segmentation

Bayesian word segmentation models (Goldwater et al., 2009)

- Learn to place boundaries after phonemes in an utterance.
- A pointwise boundary sampling algorithm: compute the probability of placing a word boundary after each phoneme.



Motivation:

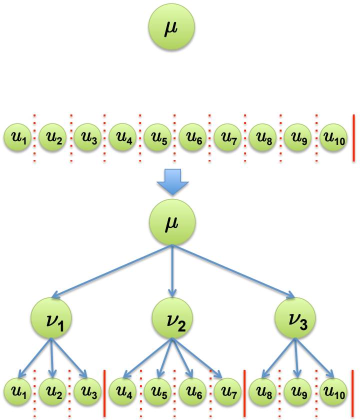
- treat text passages like phonemes in the model,
- *i.e.*, estimate text passage boundaries as per phoneme boundaries.

Outline


discovery
information retrieval
components
hierarchical multinomial
semantics
topic model
latent proportions
independent component analysis
correlations variable
Dirichlet model
nonnegative matrix factorization
variational admixture
Gibbs sampling
statistical machine learning
documents LSA
PLSI Bayesian text
natural language
unsupervised
clustering likelihood
relations estimation

Problem

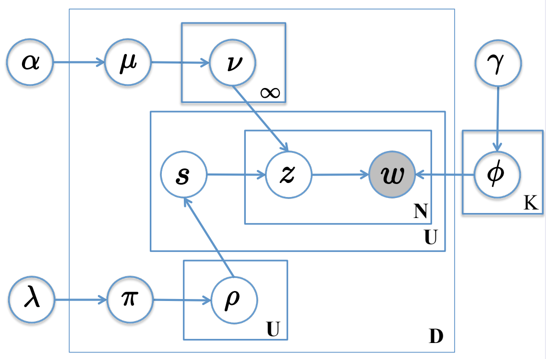
Problem:



Hypothesis: simultaneously learning topic segmentation and topic identification should allow better detection of topic boundaries.

Segmentation Model–Generative process

A segmentation model

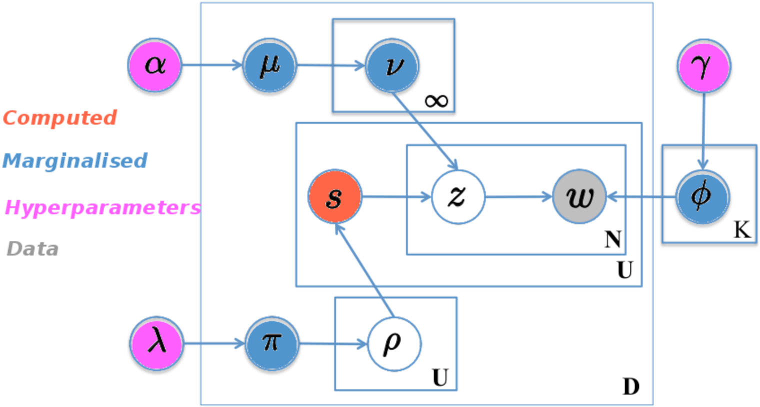


- Generative process

$$
\begin{aligned}
\vec{\phi} &\sim \text{Dirichlet}(\vec{\gamma}) \\
\vec{\mu} &\sim \text{Dirichlet}(\vec{\alpha}) \\
\pi &\sim \text{Beta}(\vec{\lambda}) \\
\vec{\nu} &\sim \text{PYP}(a, b, \vec{\mu}) \\
\rho &\sim \text{Bernoulli}(\pi) \\
z &\sim \text{Discrete}(\vec{\nu}_s) \\
w &\sim \text{Discrete}(\vec{\phi}_z)
\end{aligned}
$$

- z: topic assignment of word w;
- N: the number of words in a passage.

Posterior Inference–General Picture



- We need to sample the topic assignments z and segment boundaries ρ.

Experiments on two meeting transcripts

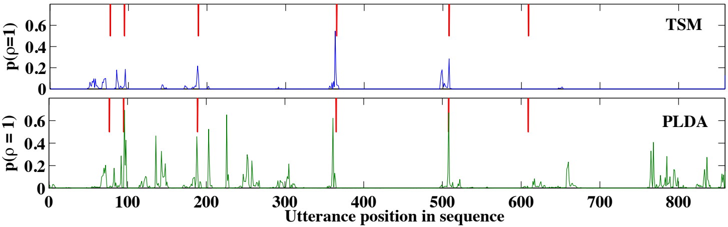


Figure: Probability of a topic boundary, compared with gold-standard segmentation on one ICSI transcript.

Gold Standard	{77,	95,	189,	365,	508,	609,	860}
PLDA	{96,	136,	203,	226,	361,	508,	860}
TSM	{85,	96,	188,	363,	499,	508,	860}

Conclusion of Segmentation with a Structured Topic Model

Paper given at NAACL 2013, main author Lan Du, also Mark Johnson

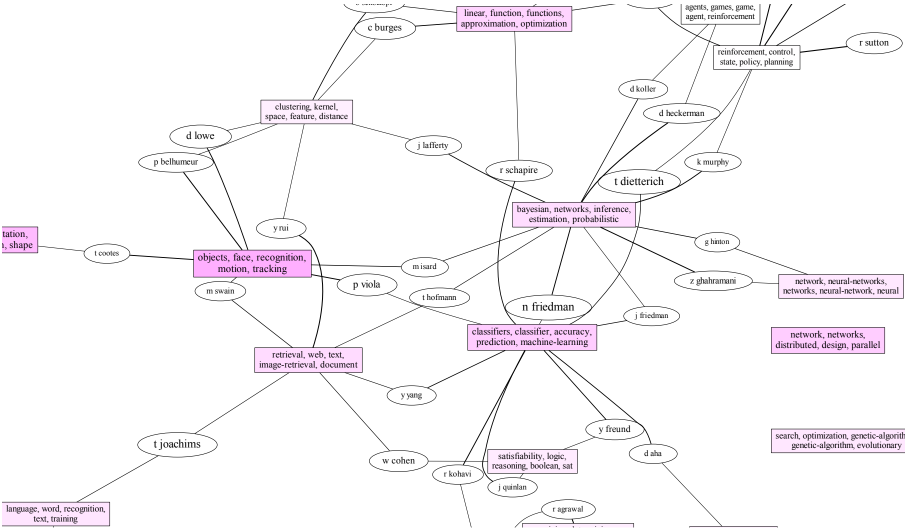
- A new hierarchical Bayesian model for unsupervised topic segmentation, using Bayesian segmentation + structured topic modelling.
- A novel sampling algorithm for splitting/merging restaurant(s) in CRP.

Conclusion of Segmentation with a Structured Topic Model

Paper given at NAACL 2013, main author Lan Du, also Mark Johnson

- A new hierarchical Bayesian model for unsupervised topic segmentation, using Bayesian segmentation + structured topic modelling.
- A novel sampling algorithm for splitting/merging restaurant(s) in CRP.
- Code is available at Lan Du's website.
- Now running multi-core.

Fun with Bibliographies (Lim *etal* ACML 2014)



Conclusion

- Latent Semantic Modelling with non-parametric Bayesians methods!
- Hierarchical stick-breaking and Chinese restaurant process methods seem inferior to block table indicator samping.
- See the individual papers.
- Read my blog and tutorials

  `https://topicmodels.org`
  "A tutorial on non-parametric methods"

Thank You ... Questions?

Alphabetic References

D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, **3**, 2003.

W.L. Buntine and S. Mishra, "Experiments with Non-parametric Topic Models," *KDD*, New York, 2014.

C. Chen, L. Du, L. and W. Buntine, "Sampling table configurations for the hierarchical Poisson-Dirichlet process," *ECML-PKDD*, Athens, 2011.

G. Doyle and C. Elkan, "Accounting for burstiness in topic models," *ICML*, Montreal, 2009.

L. Du, W. Buntine and H. Jin, "A segmented topic model based on the two-parameter Poisson-Dirichlet processh," *Machine Learning Journal*, **81**(1), 2010.

L. Du, W. Buntine and M. Johnson, "Topic segmentation with a structured topic model," *NAACL-HLT*, Atlanta, 2013.

S. Goldwater, T. Griffiths and M. Johnson, "A Bayesian Framework for Word Segmentation: Exploring the Effects of Context," *Cognition*, **112**(1), 2009.

L.F. James, "Poisson Latent Feature Calculus for Generalized Indian Buffet Processes," *arXiv*:1411.2936, November 2014.

Alphabetic References

D. Newman, J.H. Lau, K. Grieser and T. Baldwin, "Automatic Evaluation of Topic Coherence," *NAACL-HLT*, Los Angeles, 2010.

A. Puurula, "Cummulative Progress in Language Models for Information Retrieval," *ADCS*, Brisbane, 2013.

S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, **3**(4), 2009.

P. Sunehag, "Using two-stage conditional word frequency models to model word burstiness and motivating TF-IDF," *Artificial Intelligence and Statistics*, 2007.

Y.W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," *ACL*, Sydney, 2006.

Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei, "Hierarchical Dirichlet Processes," *JASA*, **101**(476), 2006.

H.M. Wallach and I. Murray and R. Salakhutdinov and D. Mimno, "Evaluation Methods for Topic Models," *ICML*, Montreal, 2009.

F. Wood, C. Archambeau, J. Gasthaus, L. James, and Y.W. Teh. "A Stochastic Memoizer for Sequence Data," *ICML*, Montreal, 2009.