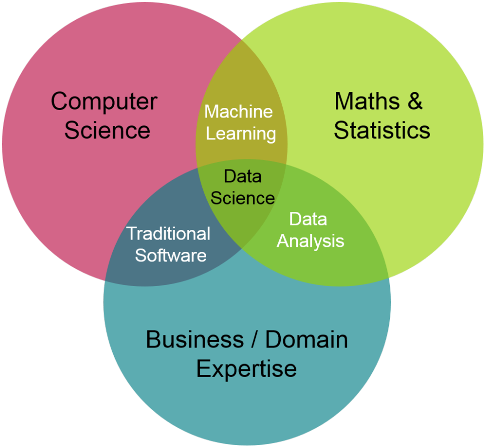
Research Career Perspective

Wray Buntine

Professor in Faculty of IT
Director of Master of Data Science
Monash University
http://topicmodels.org

2018-03-23

My Research Area



Domain expertise: semi-structured data, NLP, health informatics

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

Simple Trees

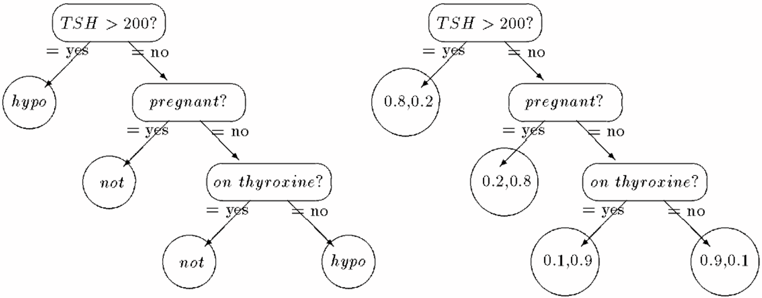
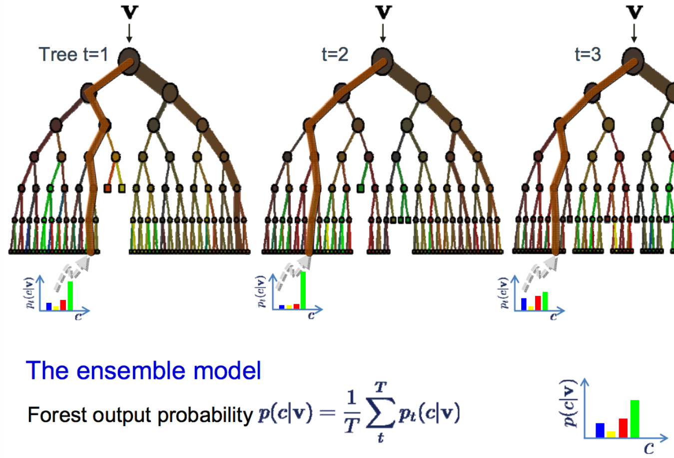


Figure 1: A decision tree and a class probability tree from the thyroid application

classification models used in 80s-90s

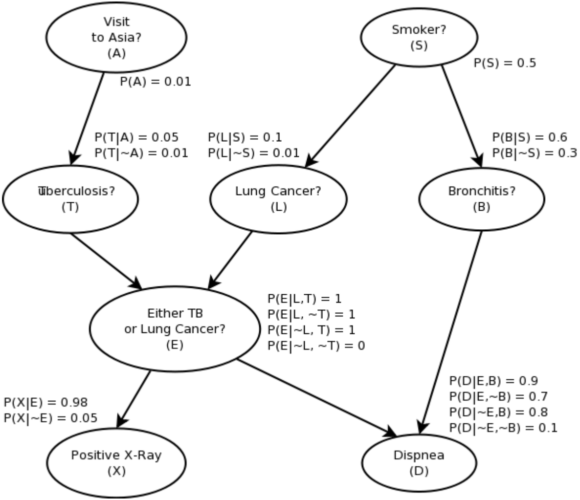
Random Forest



from *TowardsDataScience.com*

state of the art classification model from Breiman

Bayesian Network



diagnostic models used since 90s

Bayesian Network with Plate

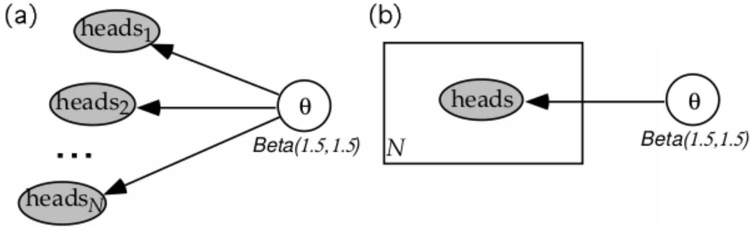


Figure 13: Tossing a coin: model without and with a plate

turn Bayesian network into a modelling language for machine learning

Linear Regression Model

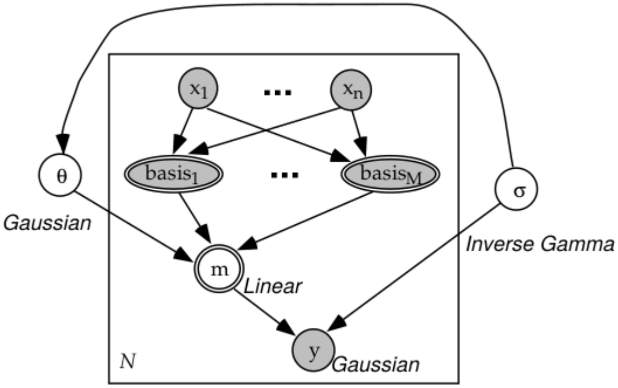


Figure 20: The linear regression problem

e.g. linear regression with non-linear basis functions and heterogeneous variance

Linear Regression, Solution

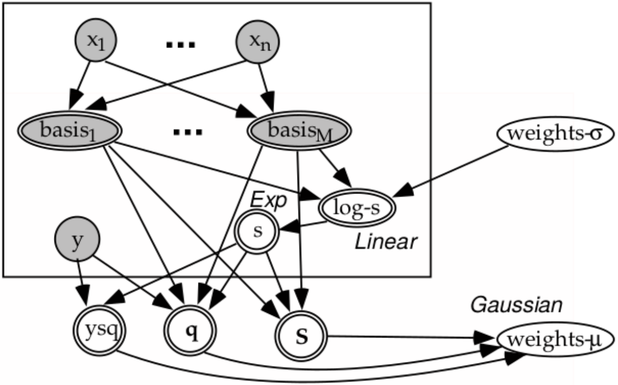


Figure 33: The heterogeneous variance problem with the plate simplified

the computation version derived from the model

Algorithms/Software for Machine Learning

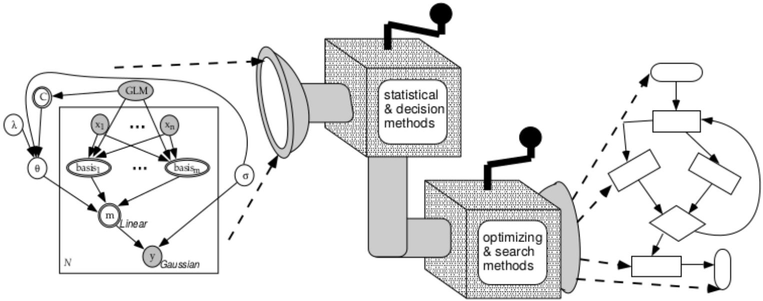
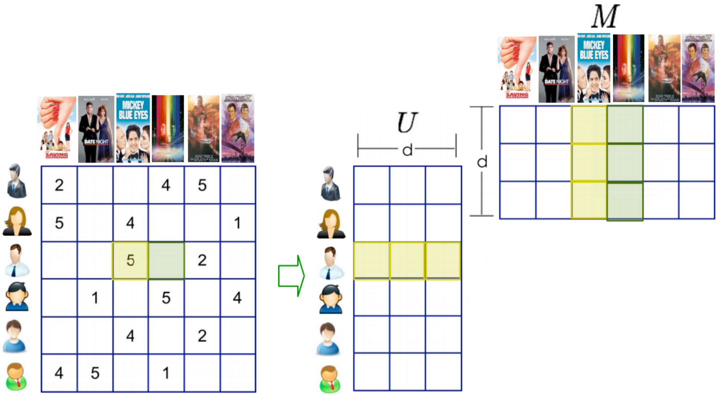


Figure 1: A software generator

my broader vision, back in 1994!

see *"Operations for Learning with Graphical Models"*, Buntine, JAIR 1994

Recommender System as Matrix Factorisation



(from Alexandros Karatzoglou, *Recommender Systems*, 2013)

- a matrix factorisation model used
- introduces latent variables called "topics" or "components"

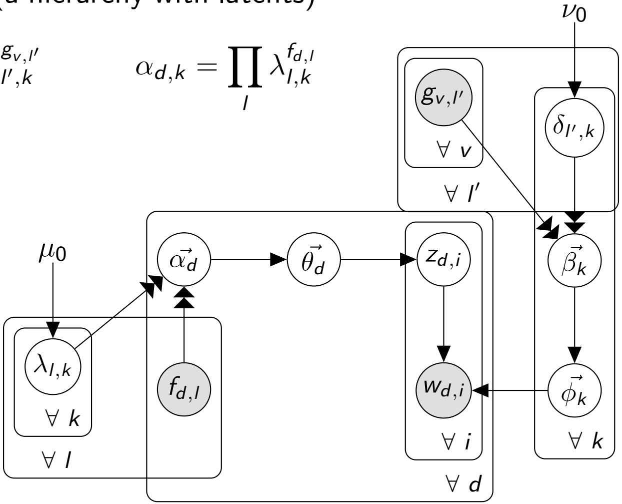
Recommender System with Side Information

- **regress from** side information
- **onto** user-item recommendations

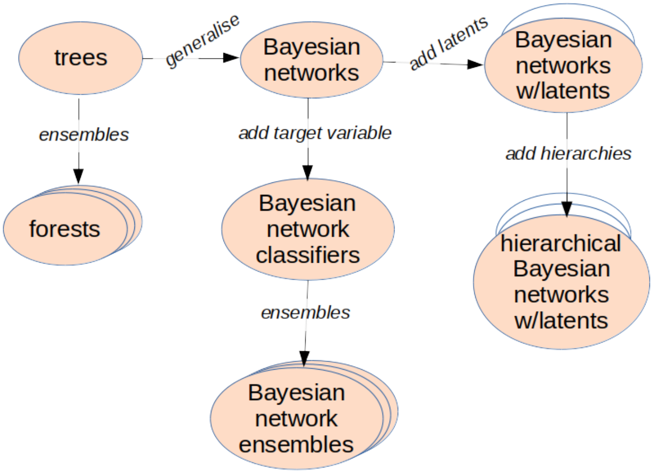
RecSys with Side-Info, model

matrix factorisation using gamma regression (a hierarchy with latents)

$$\beta_{k,v} = \prod_{l'} \delta_{l',k}^{g_{v,l'}} \qquad \alpha_{d,k} = \prod_{l} \lambda_{l,k}^{f_{d,l}}$$



Modelling Summary



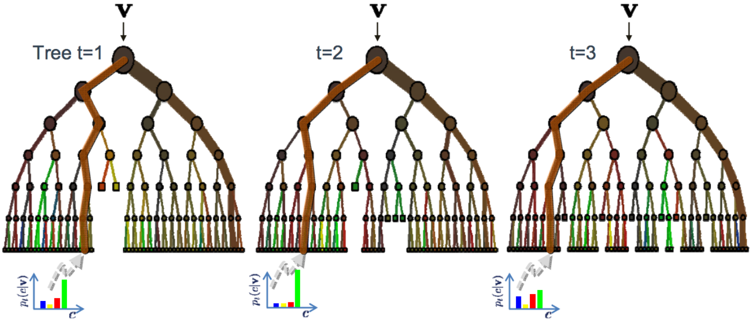
Outline


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

Historical Context

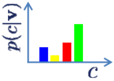
- Bayesian model averaging (BMA)
- non-parametric hierarchical models

BMA for Trees



The ensemble model

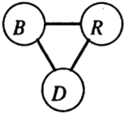
Forest output probability $p(c|\mathbf{v}) = \frac{1}{T}\sum_{t}^{T} p_t(c|\mathbf{v})$



developed early 90s, inspired Breiman's random forests

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 Storyline

1990: Graham Williams PhD thesis, ensembles for decision trees.

1990: My PhD thesis, BMA for decision trees.

1990: York and Madigan develop BMA for Bayesian networks.

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

- random forests

→ 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 counterparts bagging/ensembles became dominant paradigms.

BMA 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** by allowing content to be modelled hierarchically.

Motivation

- While somewhat successful, the BMA paradigm based on standard (current in year 2000) methods had reached a limit.
- Solution requires efficient non-parametric hierarchical modelling.
- Which we now have.

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

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.

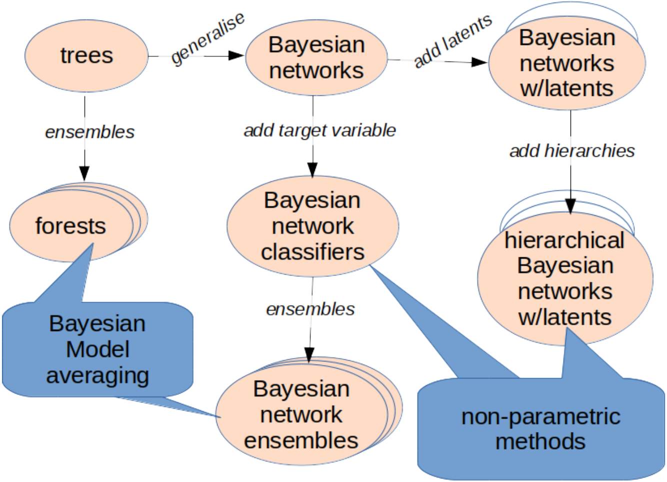
2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes (HDP), *e.g.* applied to LDA.

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 **collapsed samplers**.

Modelling Summary, cont.



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

Hierarchical Extensions to Various Models

- extend matrix factorisation, LDA and BNCs in various ways
- to develop state of the art algorithms

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
estimation
relations

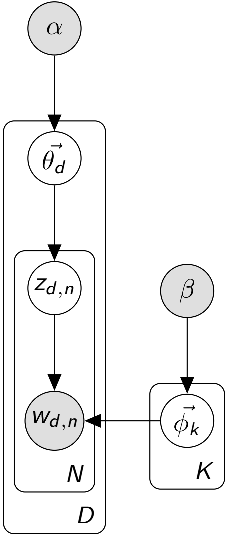
Non-Parametric Topic Model

Material from "Experiments with Non-parametric Topic Models," Buntine and Mishra, *KDD 2014*.

- Both the "word side" and the "document side" of the model could use hierarcchical priors.
- How could this be implemented efficiently?

We would like to use non-parametric methods to build better topic models.

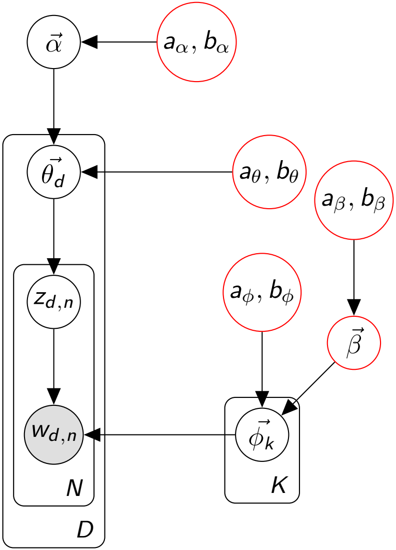
Evolution of Models: Basic LDA



LDA

α and β are concentrations;
originally parameters supplied by user;
a symmetric Dirichlet prior on rows of $\mathbf{\Theta}$ and $\mathbf{\Phi}$;
later versions use an asymmetric Dirichlet prior

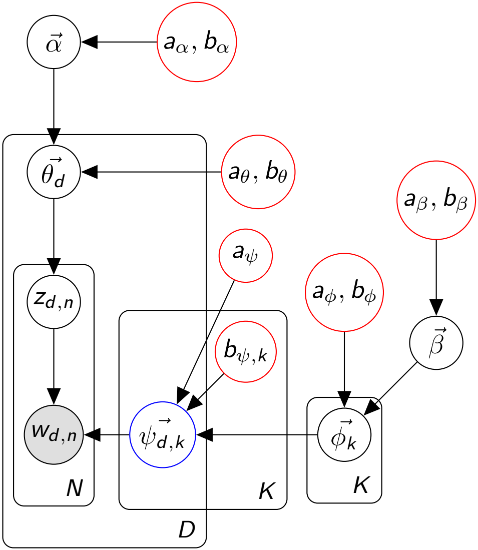
Evolution of Models: NP-LDA (2013)



NP-LDA

adds power law on word distributions like Sato and Nakagawa (2010) and estimation of background word distribution

Evolution of Models: Bursty NP-LDA (2014)

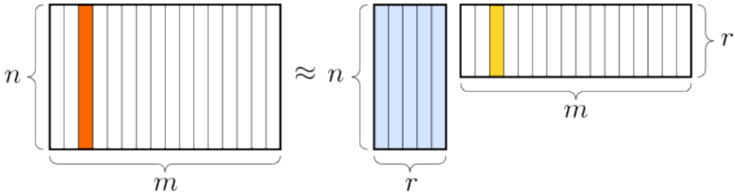


NP-LDA with Burstiness

add's burstiness like Doyle and Elkan (2009)

Example Topics

- 2691 abstracts from JMLR vol. 1-11
- about 60 words per abstract (after removing stops words)
- about 4600 words in vocabulary
- built 1000 topics
- dimensions $m = 2691$ $n = 4662$, $r = 1000$



Example Topics with “posterior”

#25, 0.42%	posteriori expectation likelihood-based analytically maximum likelihood maximization e-step posterior estimation map coming model *(top=14)*
#31, 0.40%	undirected graphical message-passing inference graphs directed posteriori junction graph clique intractable models cliques partition *(top=13)*
#38, 0.37%	dirichlet bayesian priors conjugate prior ill-posed posterior covariance gaussian infer serves distribution analytical distributions *(top=9)*
#58, 0.31%	latent variables discover posterior dependencies variable models modeling hidden parent unobserved part correlations constituent *(top=11)*
#95, 0.22%	particle kalman tracking filter filtering observer state appearance implement dynamics visual occlusion posterior multimodal *(top=5)*
#124, 0.19%	monte carlo chain markov jump mcmc chains reversible iterates mix proposal posterior problematic sampling *(top=2)*
#239, 0.11%	naive classifier bayes averaging logarithmic multiple-instance averaged posterior already counterpart considerably weakly classifications goal *(top=3)*
#678, 0.03%	nondeterministic posterior probabilities cancer hypotheses successively true deterministic distributions worst-case comprise discussed limiting *(top=2)*
#820, 0.02%	estimators constrains sparsely constraints cross-validated insufficient parameters made among posterior context brain correlated fast *(top=1)*

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

Document Segmentation

Material from "Topic segmentation with a structured topic model" Du, Buntine and Johnson, *NAACL-HLT 2013*.

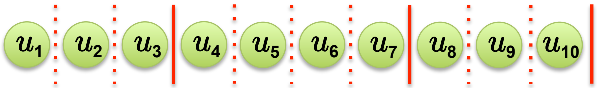
- Both the "word side" and the "document side" of the model could use hierarcchical priors.
- How could this be implemented efficiently?

We would like to use non-parametric methods to build better topic models.

Bayesian Segmentation

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

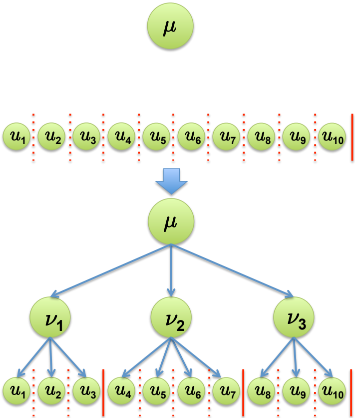
- Learn to place boundaries after phonemes in an utterance.
- A pointwise boundary sampling algorithm: compute the probability of placing a word boundary after each phoneme.



- Prediction task: predict where to place segment boundaries.

Hierarchical Bayesian Segmentation

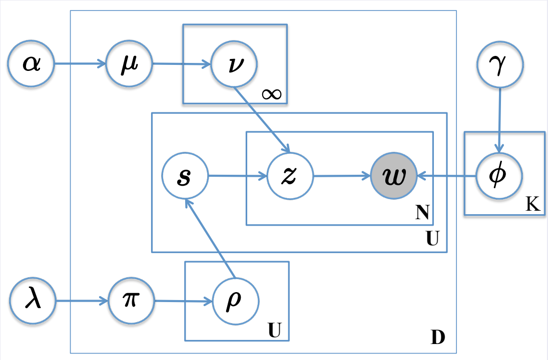
Model Concept:



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.

Experiments on two meeting transcripts

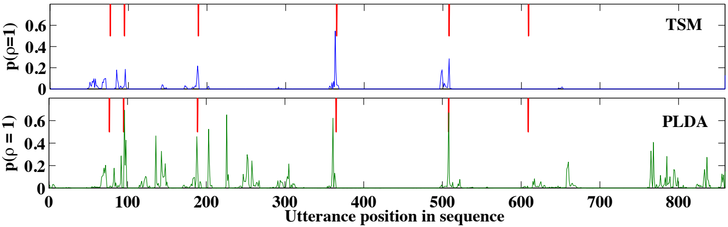


Figure: Probability of a topic boundary, compared with gold-standard segmentation on one ICSI transcript.

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}

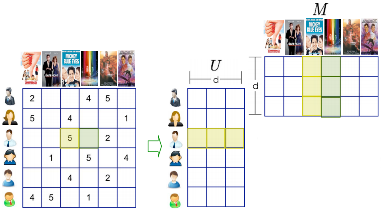
Outline

1. Models and Modelling
2. Historical Context
3. **Research Highlights**
   - Non-Parametric LDA
   - Document Segmentation
   - **Topic Models with Side Information**
   - Bayesian Network Classifiers
4. Research Agenda

Topic Model with Side Information

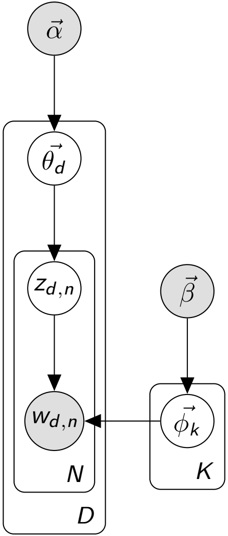
Material from "MetaLDA: a Topic Model that Efficiently Incorporates Meta information," Zhao, Du and Buntine, *ICDM 2017*.

- documents come with side info.: tags, author, etc.
- words come with side info.: WordNet, deep neural network, embeddings



We would like to use the side information to build better topics and matrix factorisations

Evolution of Models: Add Side Information to LDA



LDA

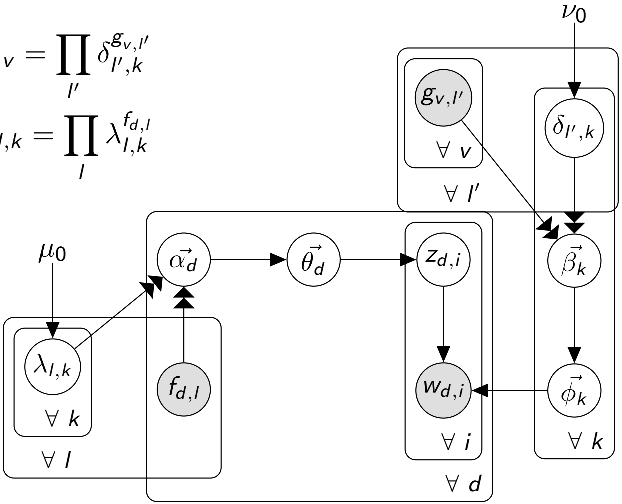
regress latent $\vec{\alpha}$ from Boolean document features $\vec{f}_d$;
regress latent $\vec{\beta}$ from Boolean word features $\vec{g}_v$;

Topic Model with Side Information

Gamma regression: topic priors $\vec{\alpha}_d$ and word priors $\vec{\beta}_k$ constructed as

$$\beta_{k,v} = \prod_{l'} \delta_{l',k}^{g_{v,l'}}$$

$$\alpha_{d,k} = \prod_{l} \lambda_{l,k}^{f_{d,l}}$$

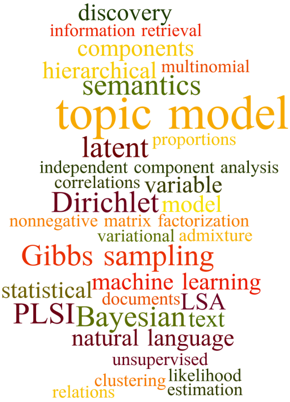


Perplexity Results on Short Text

ExtLDA = our algorithm
ExtLDA-no = our algorithm with no side information
(variant of earlier NP-LDA)
DMR = previous state of the art algorithm using logistic regression

	Web Snippets corpus				Tag My News corpus			
	50	*100*	*150*	*200*	*50*	*100*	*150*	*200*
LDA	961	878	869	888	1969	1873	1881	1916
ExtLDA	**774**	**627**	**572**	**534**	**1657**	**1415**	**1304**	**1235**
ExtLDA-no	884	733	671	625	1800	1578	1469	1422
DMR	845	683	607	562	1750	1506	1391	1323
LF-LDA	1162	1076	1016	1012	2436	2404	2394	2396
WF-LDA	894	839	827	842	1853	1766	1830	1854
LLDA	1543				2958			
PLLDA	*5*	*10*	*20*	*50*	*5*	*10*	*20*	*50*
	1060	886	735	642	2181	1863	1647	1456

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

Simple BNCs

Prediction task: predict Y given $X_1, ..., X_4$

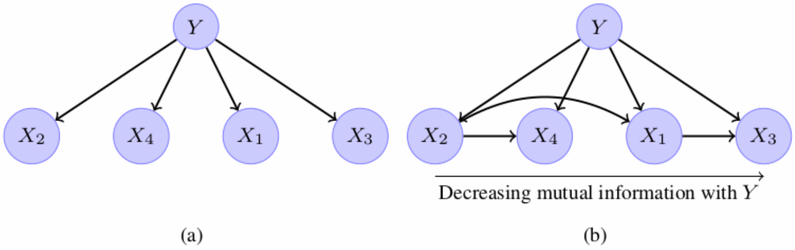


Fig. 1 Example BNC structures: (a) Naïve Bayes, (b) kDB-1

Bayesian Network Classifiers make the target variable the root of the network. More parents means more complex model, and sometimes better classification.

BNCs with Hierarchical Priors

- Add Dirichlet process priors to the class probability tables of the BNC.
- Yields far better probability estimates during learning
- Substantially beats existing methods for BNCs

- Without ensembles: better than Random Forests.
- Scales on larger data, comparable to XGBoost (another tree algorithm).
- With ensembles: substantially better than Random Forests.

to appear in "Accurate parameter estimation for Bayesian network classifiers using hierarchical Dirichlet processes," Peptitjean , Buntine, Webb and Zaidi, ECML-PKDD 2018

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

Motivation

- Deep neural networks (DNNs) are broadly successful (partially) because of ease of implementation.
- Probability networks allow broad range of models complementing DNNs.
- Statistic processes particular are more challenging for the "non-initiates" to use and implement.
- We have realised our techniques can be automated!

Research Goal

Goal

How to build support tools, like those used for deep neural networks, for probability network models on semi-structured data?

So we would like:

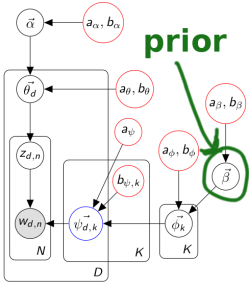
1. ability to turn network specification into code,
2. flexible range of models supported,
3. state of the art algorithms,
4. compilation down to CPU clusters or GPUs.

- Our group is currently exploring the items 1-3.
- We're also exploring the better neural net methods.

Questions?

Questions?

Understanding the Word Prior $\vec{\beta}$

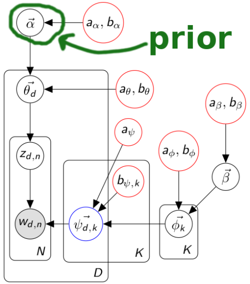


- word prior $\vec{\beta}$ corresponds to a background topic (with PYPs to make Zipfian)
- all topics $\vec{\phi}_k$ are variants of it
- "topical" words are reduced and "stop" words are increased compared to collection frequency
- word w importance for topic k can be measured as $\phi_{k,w}/\beta_w$

395 Reuters RCV1 news articles from 1996 containing "church".

$\vec{\beta}$ background (high β_w)	the of to a in and 's was on for by telephone said sick fighting with as at is republican land shortly he remains difficult voice shown mary inside done travelled
topical (high df_w/β_w)	diana teresa missionaries russia parker elizabeth bowles camilla churchill winston harriman pamela her quoted princess pontiff prince navarro-valls dies kremlin averell parkinson
topic #1	'm else something n't everyone someone my stand i me like truth always really going do you ran know similar lover things look sun think 've

Understanding the Topic Prior $\vec{\alpha}$

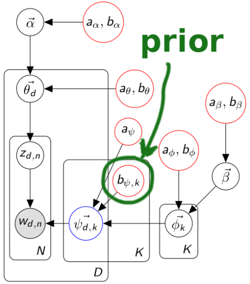


- topic prior $\vec{\alpha}$ corresponds to a expected occurence of topic
- all document topics $\vec{\theta}_d$ are variants of it
- uses Dirichlet processes

395 Reuters RCV1 news articles from 1996 containing "church".

#2 5.05%	quoted declined saying reports position further talks sought	#47 1.00%	nazi nazis germany german war jews recalled forces wartime
#4 2.28%	pope vatican navarro-valls pontiff solidarity 76-year-old	#48 1.85%	estimated sold york company estate island percent sell money
#5 1.90%	diana charles bowles parker prince camilla princess elizabeth	#49 0.98%	art works culture includes cultural st boris programme show

Understanding the Topic Confidence $b_{\psi,k}$



- topic confidence $b_{\psi,k}$ corresponds to a concentration parameter (inverse variance) when generating $\vec{\psi}_{d,k}$ from $\vec{\phi}_k$
- large values means $\vec{\psi}_{d,k}$ is a copy of $\vec{\phi}_k$
- low values means $\vec{\psi}_{d,k}$ differs greatly from $\vec{\phi}_k$
  - if really low, its a “rubbish” topic

8616 Reuters RCV1 news articles from 1996 containing “person”.

#2 2605	willingness guarantor caution definite absences disgruntled seriousness maneouvring govern instability
#4 2271	detractors predecessors illustrious front-runner outsider credentials flair courteous woo self-effacing married
#9 2153	teresa missionaries woodlands birla pacemaker gutters nun calcutta

#199 3.57	royalties michelle job lopez earns hungarian-born eating credits
#200 2.92	penguin birthdays 1000 abc compiled wheel mausoleum 1800 provoke timetable budapest
#194 1.41	spa verdicts korzhakov beginnings burmese ethics betrayed blair fujimori heroin

General Methods on Networks: Summary

- There is a rich variety of methods for doing inference and learning on complex probabilistiic networks.
- Making it more automatic is our challenge.
- Existing software from statistics indicates feasibility:
  - *BUGS*
  - *Just Another Gibbs Sampler (JAGS)*
  - *Stan*
- Existing DNN software suggests implementation ideas.

I am writing a text-book targeted at 3rd year data science students on **probability networks for machine learning**.