Variational Bayes for generic topic models

Gregor Heinrich and Michael Goesele

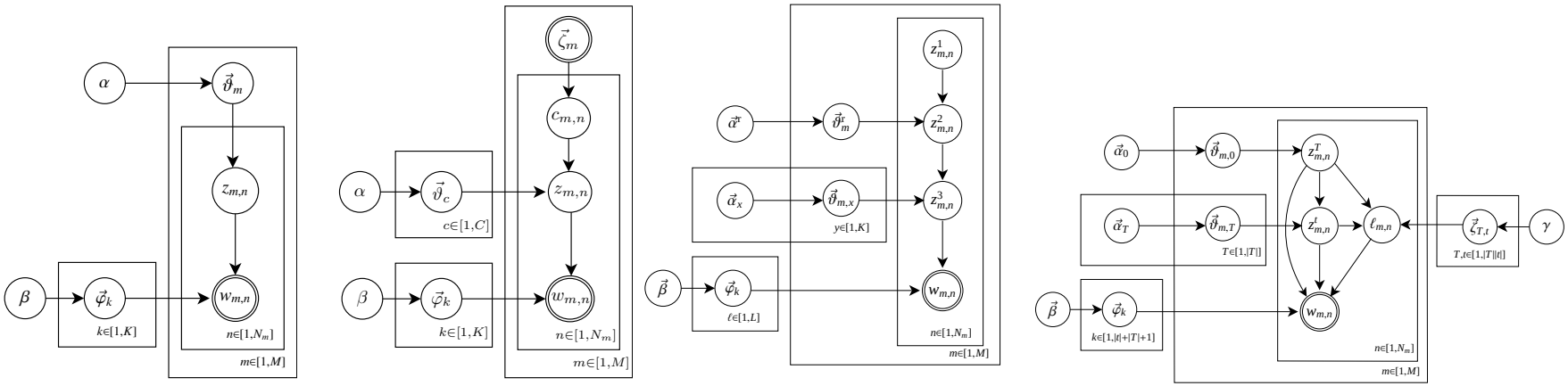
[1]Fraunhofer IGD + University of Leipzig; [2]TU Darmstadt

16 September 2009

28th Annual Conference on Artificial Intelligence

KI2009

Topic models



(a) LDA (b) ATM (c) PAM (d) hPAM

- Probabilistic models for co-occurrences in discrete data
- Close relationship to principal components analysis (PCA) and latent semantic analysis (LSA)
- Generalise latent Dirichlet allocation (LDA) (Blei et al. 2003) to accommodate structures assumed in data
- Used in many areas of data-driven artificial intelligence:
  - Text mining: retrieval, automatic thesaurus generation;
  - Data mining: social networks, authorship analysis;
  - Computer vision: image classification and content based retrieval;
  - etc.

Example: text mining

Topic label	Most likely words
political parties	CDU Partei Kohl Aufklärung Schäuble Zeitung Union Krise Wahrheit Affäre Christdemokraten Glaubwürdigkeit Konsequenzen
soccer premier ligue	FC SC München Borussia SV VfL Kickers SpVgg Uhr Köln Bochum Freiburg VfB Eintracht Bayern Hamburger Bayern+München
police / accident	Polizei verletzt schwer Auto Unfall Fahrer Angaben schwer+verletzt Menschen Wagen Verletzungen Lawine Mann vier Meter Straße
Chechnya	Rebellen russischen Grosny russische Tschetschenien Truppen Kaukasus Moskau Angaben Interfax tschetschenischen Agentur
politics / Hesse	FDP Koch Hessen CDU Koalition Gerhardt Wagner Liberalen hessischen Westerwelle Wolfgang Roland+Koch Wolfgang+Gerhardt
weather	Grad Temperaturen Regen Schnee Süden Norden Sonne Wetter Wolken Deutschland zwischen Nacht Wetterdienst Wind
politics / Croatia	Parlament Partei Stimmen Mehrheit Wahlen Wahl Opposition Kroatien Präsident Parlamentswahlen Mesic Abstimmung HDZ
green party	Grünen Parteitag Atomausstieg Trittin Grüne Partei Trennung Mandat Ausstieg Amt Roestel Jahren Müller Radcke Koalition
Russian politics	Russland Putin Moskau russischen russische Jelzin Wladimir Tschetschenien Russlands Wladimir+Putin Kreml Boris Präsidenten
police / schools	Polizei Schulen Schüler Täter Polizisten Schule Tat Lehrer erschossen Beamten Mann Polizist Beamte verletzt Waffe

LDA topics, 18400 DPA news articles, Jan. 2000 (Heinrich et al. 2005)

Overview

Context:

- Motivation: For topic models beyond latent Dirichlet allocation, mostly Gibbs sampling has been used as inference scheme
- Goal: Explore variational Bayes for topic models in general (rather than specific for some given model)
- Complement generic Gibbs sampling (Heinrich 2009, ECML)

Structure:

- LDA and generalisations
- Generic approach to Variational Bayes
- Experiments
- Conclusions, ongoing and future work

Overview

Context:

- Motivation: For topic models beyond latent Dirichlet allocation, mostly Gibbs sampling has been used as inference scheme
- Goal: Explore variational Bayes for topic models in general (rather than specific for some given model)
- Complement generic Gibbs sampling (Heinrich 2009, ECML)

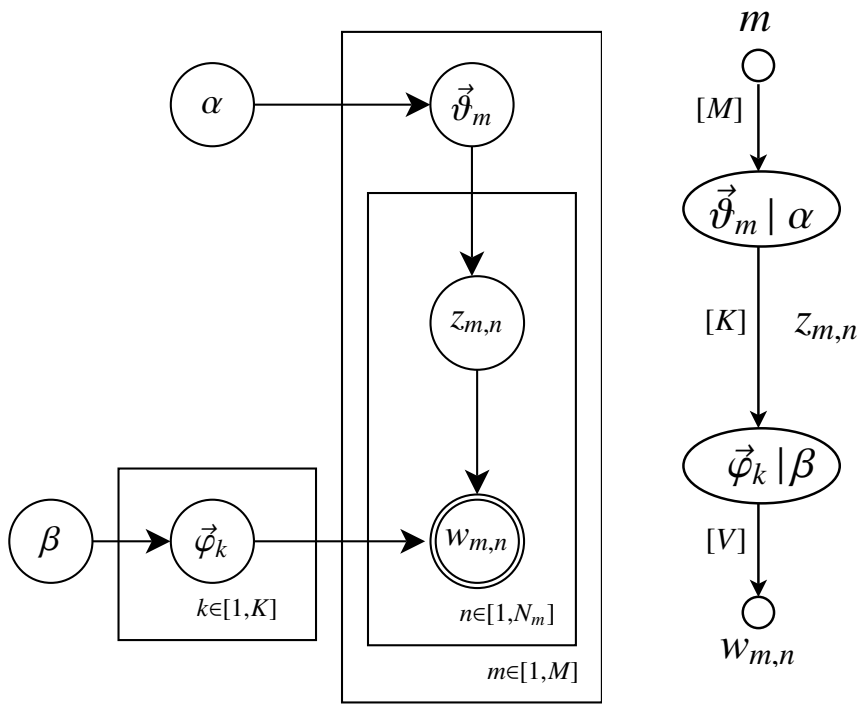
Structure:

- LDA and generalisations
- Generic approach to Variational Bayes
- Experiments
- Conclusions, ongoing and future work

Latent Dirichlet allocation

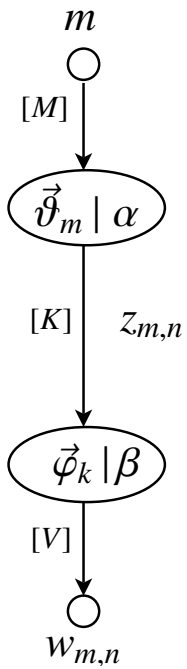
- Mixture model: complex distribution via sum of component distributions, $p(x) = \sum_{k=1}^{K} p(x|z{=}k)p(z{=}k) = \sum_{\text{dim}} \text{component} \times \text{weight}$.
- Two coupled mixtures:
  - Document m = mixture of topics z with components $\vec{\vartheta}_m = p(z|m)$ and
  - Topic z = mixture of words w with components $\vec{\varphi}_k = p(w|z{=}k)$ and component weights $\vec{\vartheta}_m$,
  - Distribution over words w, $p(w|m) = \sum_{k=1}^{K} \varphi_{k,w}\, \vartheta_{m,k}$.

```
// document plate:
for all documents m ∈ [1, M] do
    sample mixture proportion ϑ⃗_m ~ Dir(α⃗)
    sample document length N_m ~ Poiss(ξ)
    // word plate:
    for all words n ∈ [1, N_m] in doc m do
        sample topic index z_{m,n} ~ Mult(ϑ⃗_m)
        sample term for word
        w_{m,n} ~ Mult(φ⃗_{z_{m,n}})
// topic plate:
for all topics k ∈ [1, K] do
    sample mixture components φ⃗_k ~ Dir(β⃗)
```

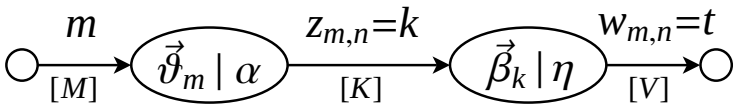


Generalisations of LDA

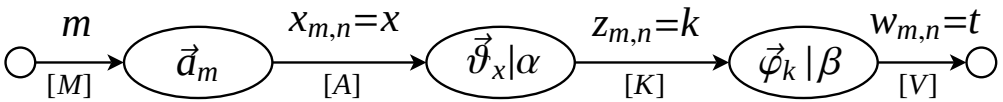
- Generic topic model characteristics:
  - Structured into "levels" with Dirichlet parameters (in LDA: φ and ϑ) and multinomial variables (in LDA: z)
  - Levels coupled via the values of discrete variables (in LDA: ϑ and φ via z).
- $\rightarrow$ Mixture networks (MNs): digraphs $G(\mathcal{N}, \mathcal{E})$
  - Node $N \in \mathcal{N}$ samples from a mixture component chosen as function of incoming edges
  - Edge $E \in \mathcal{E}$ propagates discrete values from parent to child node.
- MNs vs. BNs:
  - MNs focus on the interrelations between discrete mixtures
  - BNs focus on dependencies between random variables and express the repetitions of data points (plate notation).



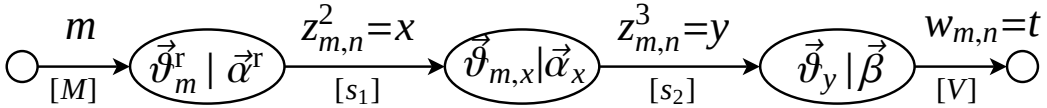
Topic models: mixture levels



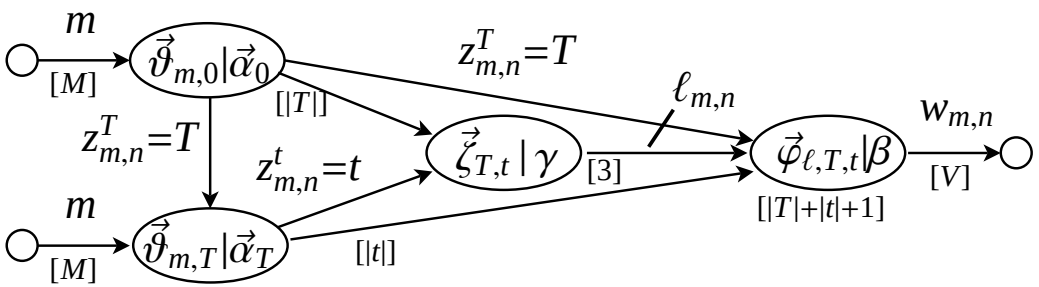
(a) Latent Dirichlet allocation, LDA



(b) Author–topic model, ATM

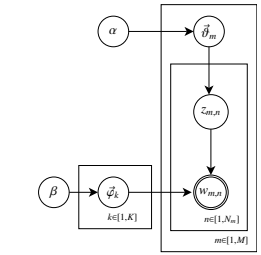


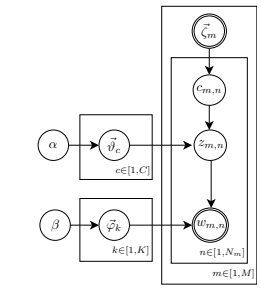
(c) Pachinko allocation model, PAM

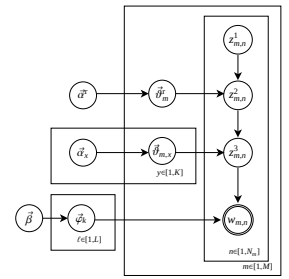


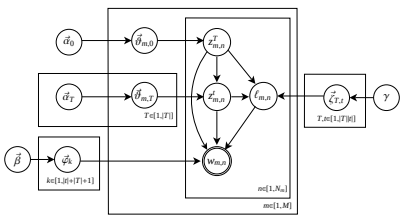
(d) Hierarchical pachinko allocation model, hPAM

(Blei et al. 2003; Rosen-Zvi et al. 2004; Li and McCallum 2006; Li et al. 2007)









Overview

Context:

- Motivation: For topic models beyond latent Dirichlet allocation, mostly Gibbs sampling has been used as inference scheme
- Goal: Explore variational Bayes for topic models in general (rather than specific for some given model)
- Complement generic Gibbs sampling (Heinrich 2009, ECML)

Structure:

- LDA and generalisations
- Generic approach to Variational Bayes
- Experiments
- Conclusions, ongoing and future work

Variational Bayes

- Posterior distribution:

$$p(H, \Theta|V) = \frac{p(V, H, \Theta)}{p(V)} = \frac{p(V, H, \Theta)}{\sum_H \int p(V, H, \Theta)\, \mathrm{d}\Theta} \,. \tag{1}$$

- Intractable $\rightarrow$ Variational Bayes (Beal 2003): relax structure of $p(H, \Theta|V)$ by simpler variational distribution $q(H, \Theta|\Psi, \Xi)$ with variational parameters Ψ and Ξ to be estimated.
- VB = bound maximisation:

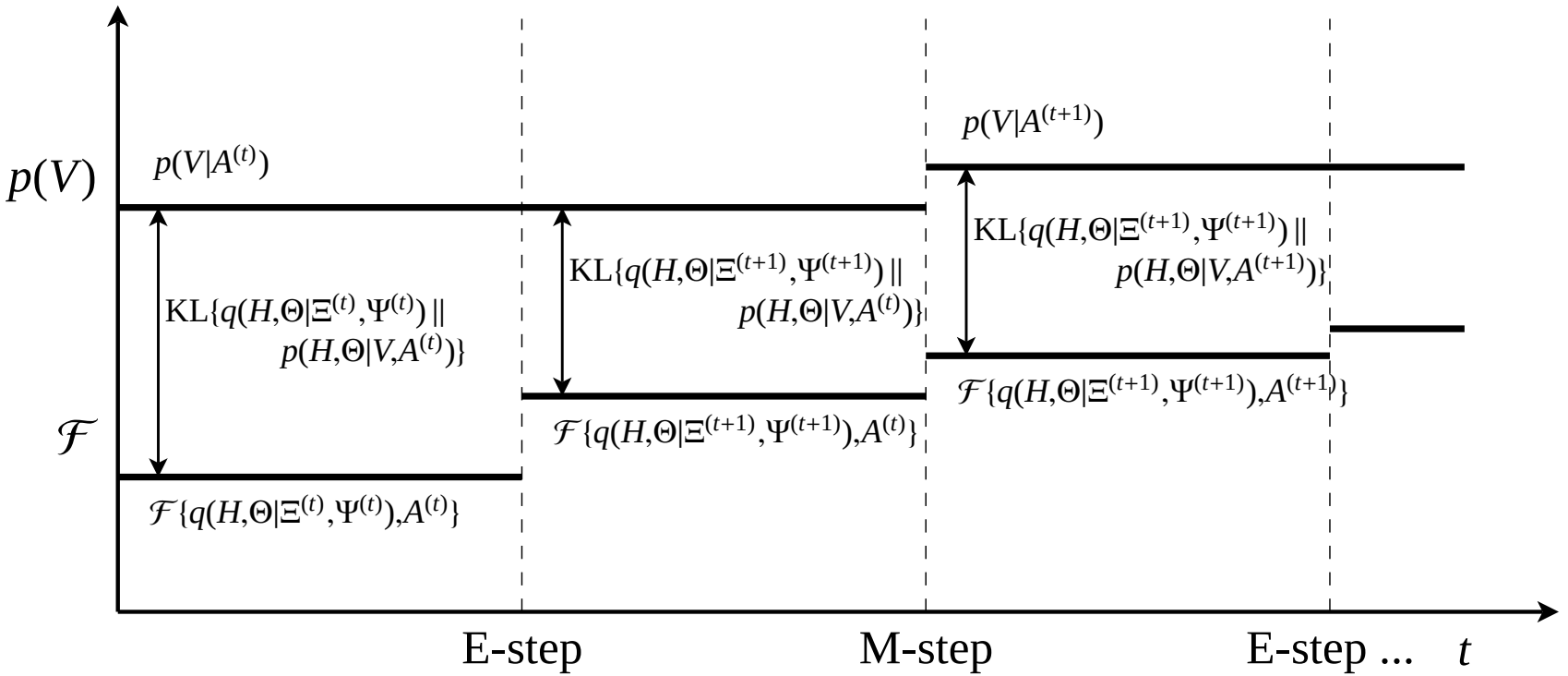
$$\log p(V) \geq \log p(V) - \mathrm{KL}\{q(H, \Theta) \parallel p(H, \Theta|V)\} \triangleq \mathcal{F}\{q(H, \Theta)\} \tag{2}$$

with free energy $\mathcal{F}\{q(H, \Theta)\}$.

Variational Bayes

Inference task: optimise $\mathcal{F}$ by EM-like algorithm:

1. E-step: max $\mathcal{F}$ w.r.t. the variational parameters
2. M-step: max $\mathcal{F}$ w.r.t. the true parameters/hyperparameters



$$\log p(V) \geq \log p(V) - \mathrm{KL}\{q(H,\Theta) \,\|\, p(H,\Theta|V)\} \triangleq \mathcal{F}\{q(H,\Theta)\} \qquad (2)$$

Mean-field approximation

- LDA: variational mean-field approach: q fully factorised:

$$q(\vec{z}, \varphi, \vartheta | \psi, \lambda, \gamma) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathrm{Mult}(z_{m,n} | \vec{\psi}_{m,n}) \cdot \prod_{k=1}^{K} \mathrm{Dir}(\vec{\varphi}_k | \vec{\lambda}_k) \prod_{m=1}^{M} \mathrm{Dir}(\vec{\vartheta}_m | \vec{\gamma}_m) \ . \tag{3}$$

- Proposal: Eq. 3 = special case of generic q that captures dependencies $\uparrow X$ between multiple hidden mixture levels:

$$q(H, \Theta | \Psi, \Xi) = \prod_{\ell \in H} \left[\prod_i \mathrm{Mult}(x_i | \underline{\psi}_i, \uparrow x_i) \right]^{[\ell]} \prod_{\ell \in L} \left[\prod_k \mathrm{Dir}(\vec{\vartheta}_k | \vec{\xi}_k, \uparrow X) \right]^{[\ell]} , \tag{4}$$

  where:
  - $\ell \in H$ refers to all levels that produce hidden variables (topics) and
  - $\underline{\psi}_i$ = joint distribution of topics over all $\ell \in H$ ("topic field")
- Eq. 4 includes LDA for the case of one hidden level ($H = \{\vec{z}\}$)

Variational E-steps

- Estimate variational distributions for the joint multinomial $\underline{\psi}_u$ for each token ("topic field") and Dirichlet parameters $\vec{\xi}^{\ell}_k$ on each level:

$$\psi_{u,\vec{t}} \propto \exp\left(\sum_{\ell \in L} \left[\mu_t(\vec{\xi}_k)\right]^{[\ell]}\right), \tag{5}$$

$$\xi^{\ell}_{k,t} = \left[\left(\sum_u n_u \psi_{u,k,t}\right) + \alpha_{j,t}\right]^{[\ell]} \tag{6}$$

- $u = (m, v)$ term tokens: join identical word tokens $i = (m, n)$.
- $\psi_{u,\vec{t}} = q(\vec{x}_u{=}\vec{t} \,|\, \underline{\psi}_u)$ = likelihood of a particular configuration of topics $\vec{t}$.
- $\psi^{\ell}_{u,k,t} = \sum_{\text{all } \vec{t} \text{ including } (k^{\ell}, t^{\ell})} \psi_{u,\vec{t}}$ = likelihood of pair (k^{ℓ}, t^{ℓ}).
- $\sum_u n_u \psi^{\ell}_{u,k,t}$ can be interpreted as the expected counts $\langle n^{\ell}_{k,t} \rangle_q$ of co-occurrence of the value pair (k^{ℓ}, t^{ℓ}).
- $\mu_t(\vec{\xi}_k) = \langle \log \vartheta_{k,t} \rangle_q$ variational expectation of parameters.

M-steps

- Estimate Dirichlet hyperparameters $\vec{\alpha}_j^{\ell}$ (or scalar α^{ℓ}) from variational expectations of model parameters $\langle \log \vartheta_{k,t} \rangle_q = \mu_t(\vec{\xi}_k)$.
- ML estimator via Newton's method (Blei et al. 2003; Minka 2000) on Dirichlet log likelihood of each mixture level.

Variants

- So far, Dirichlet priors on all multinomials.
- ML point estimates (e.g., unsmoothed LDA (Blei et al. 2003), generally levels without document-specific components).
- Modified E-step, additional M-step equation:

$$\psi_{u,\vec{t}} \propto \exp\left(\sum_{\ell \in L \setminus c} \left[\mu_t(\vec{\xi}_k)\right]^{[\ell]}\right) \cdot \vartheta^c_{k,t}\,, \tag{7}$$

$$\vartheta^c_{k,t} = \frac{\langle n_{k,t}\rangle_q}{\langle n_k\rangle_q} \propto \sum_u n_u \psi^c_{u,k,t}\ . \tag{8}$$

- Eq. 7 also for observed parameters (representing labels).

Overview

Context:

- Motivation: For topic models beyond latent Dirichlet allocation, mostly Gibbs sampling has been used as inference scheme
- Goal: Explore variational Bayes for topic models in general (rather than specific for some given model)
- Complement generic Gibbs sampling (Heinrich 2009, ECML)

Structure:

- LDA and generalisations
- Generic approach to Variational Bayes
- Experiments
- Conclusions, ongoing and future work

Experiments: Setting

- Models: LDA, ATM and PAM, unsmoothed (ML) + smoothed VB
- Baseline: Gibbs sampling implementations
- Criteria:
  1. Ability to generalise to test data V' given the model parameters Θ.
  2. Convergence time in single thread.
- Perplexity to measure generalisation to test data.
- NIPS corpus: $M = 1740$ documents (174 held-out), $V = 13649$ terms, $W = 2301375$ tokens, $A = 2037$ authors.

Experiments: Results

Model:		LDA			ATM			PAM		
Dimensions {A,B}:		$K = \{25, 100\}$			$K = \{25, 100\}$			$s_{1,2} = \{(5, 10), (25, 25)\}$		
Method:		GS	VB_{ML}	VB	GS	VB_{ML}	VB	GS	VB_{ML}	VB
Convergence time [h]	A	0.39	0.83	0.91	0.73	1.62	1.79	0.5	1.25	1.27
	B	1.92	3.75	4.29	3.66	7.59	8.1	5.61	14.86	16.06
Iteration time [sec]	A	4.01	157.3	164.2	6.89	254.3	257.8	5.44	205.1	207.9
	B	16.11	643.3	671.0	29.95	1139.2	1166.9	53.15	2058.2	2065.1
Iterations	A	350	19	20	380	23	25	330	22	22
	B	430	21	23	440	24	25	380	26	28
Perplexity	A	1787.7	1918.5	1906.0	1860.4	1935.2	1922.8	2053.8	2103.0	2115.1
	B	1613.9	1677.6	1660.2	1630.6	1704.0	1701.9	1909.2	1980.5	1972.6

- VB: perplexity in the range of Gibbs counterparts
- Full VB slightly better than ML.
- VB consistently weaker to Gibbs baseline (adverse initialisation, correlation between Ψ and Ξ assumed independent?)
- VB implementations half as fast as Gibbs samplers.

Overview

Context:

- Motivation: For topic models beyond latent Dirichlet allocation, mostly Gibbs sampling has been used as inference scheme
- Goal: Explore variational Bayes for topic models in general (rather than specific for some given model)
- Complement generic Gibbs sampling (Heinrich 2009, ECML)

Structure:

- LDA and generalisations
- Generic approach to Variational Bayes
- Experiments
- Conclusions, ongoing and future work

Conclusions

- Variational Bayes derivation for a large class of topic models by generalising LDA.
- Algorithm that can be easily applied to specific topic models.
- Application to example models, verifying the general applicability.
- So far, especially more complex topic models have predominantly used inference based on Gibbs sampling
- This paper = step towards exploring the possibility of variational approaches: Unifies theory of topic models in general including labels, point estimates and component grouping for VB.
- More work remains to be done in order to make VB algorithms as effective and efficient as their Gibbs counterparts.

Ongoing and future work

- Optimise implementations → improve the experimental results
- Develop parallel algorithms
- Extend framework of generic topic models
- Harmonise with collapsed Gibbs sampling for generic topic models (Heinrich 2009)
- Generalisation of collapsed variational Bayes (Teh et al. 2007).

Questions ?

References I

References

Beal, M. J. (2003).
Variational Algorithms for Approximate Bayesian Inference.
Ph. D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Blei, D., A. Ng, and M. Jordan (2003, January).
Latent Dirichlet allocation.
Journal of Machine Learning Research 3, 993–1022.

Heinrich, G. (2009).
A generic approach to topic models.
In *Proc. European Conf. on Mach. Learn. / Principles and Pract. of Know. Discov. in Databases (ECML/PKDD) (in press)*.

Heinrich, G., J. Kindermann, C. Lauth, G. Paaß, and J. Sanchez-Monzon (2005).
Investigating word correlation at different scopes—a latent concept approach.
In *Workshop Lexical Ontology Learning at Int. Conf. Mach. Learning*.

Li, W., D. Blei, and A. McCallum (2007).
Mixtures of hierarchical topics with pachinko allocation.
In *International Conference on Machine Learning*.

References II

Li, W. and A. McCallum (2006).
Pachinko allocation: DAG-structured mixture models of topic correlations.
In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, New York, NY, USA, pp. 577–584. ACM.

Minka, T. (2000).
Estimating a Dirichlet distribution.
Web.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth (2004).
The author-topic model for authors and documents.
In *20th Conference on Uncertainty in Artificial Intelligence*.

Teh, Y. W., D. Newman, and M. Welling (2007).
A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation.
In *Advances in Neural Information Processing Systems*, Volume 19.