A generic approach to topic models

Gregor Heinrich
CTO, vsonix GmbH, Darmstadt
http://vsonix.com
http://arbylon.net

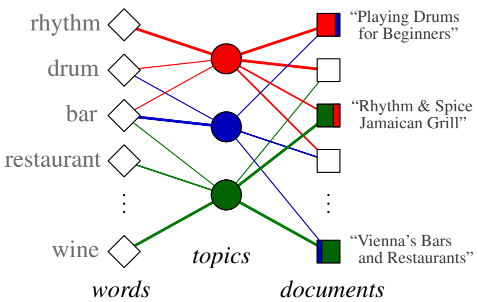
Research Seminar, Institute for Statistics and Mathematics
Vienna University of Economics and Business
http://www.wu.ac.at/statmath/resseminar

Vienna, 1 June 2012

Overview

- Topic models – motivation and review
- Networks of mixed membership (NoMMs)
- Inference – a Gibbs "meta-sampler"
- NoMM typology and design
- Application to tag-enhanced expertise finding
- Conclusions and outlook

Topic models



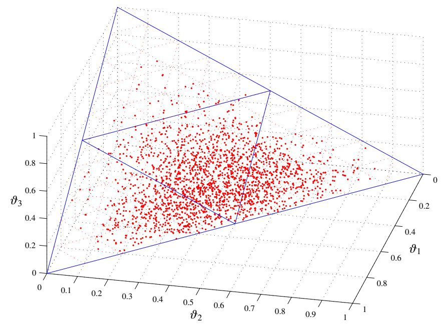
- Probabilistic representations of grouped discrete data
- Illustrative for text: words grouped in documents
  - Latent topics (a.k.a. concepts, components) = cluster semantically related words (Landauer and Dumais 1997; Griffiths et al. 2007)
  - Language = semantic meaning (topics) + noise

→ Reduce vocabulary problem by discovery of semantic relations
→ Reduce sparsity problem by dimensionality reduction ↔ discrete principal components analysis (Buntine and Jakulin 2005)

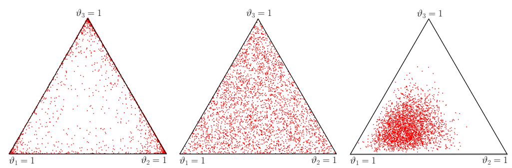
Towards Bayesian topic models: the Dirichlet distribution

Bayesian methodology:

- Parameters generated from *prior* distributions
- Language data: popular prior for the multinomial / discrete distribution: Dirichlet distribution
  - Conjugacy: straight-forward mathematical form
- Bayesian topic model: Latent Dirichlet allocation (Blei et al. 2003)

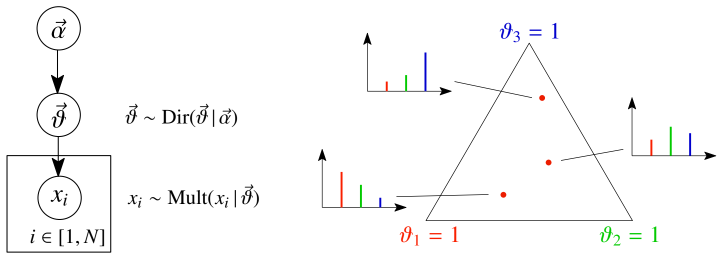


(a) Dir(4, 4, 2)



(b) Dir(.1, .1, .2) (c) Dir(1, 1, 1) (d) Dir(8, 4, 3)

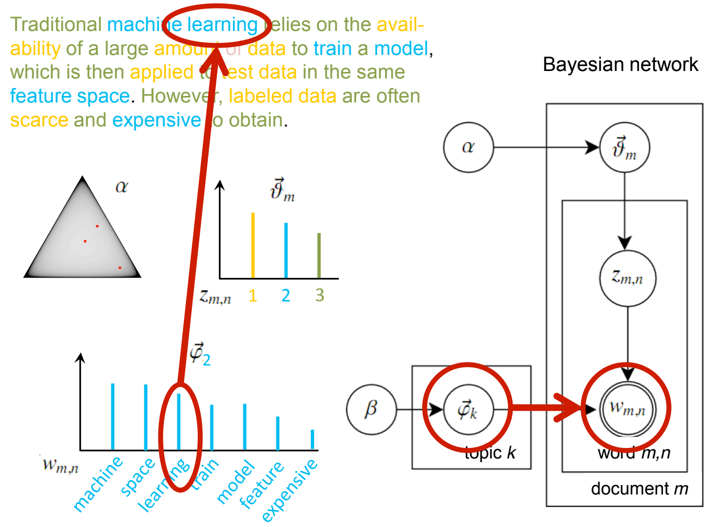
Bayesian networks: Dirichlet-generated multinomials



Bayesian networks:

- Graphical modelling of joint probability distributions
- Node: random variable
- Edge: conditional probability distribution
- Plate: repeated i.i.d. samples

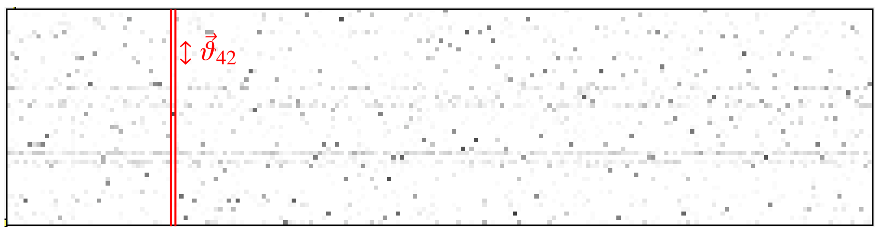
Latent Dirichlet Allocation



Draw word from term distribution of topic 2, "learning"

Example document–topic distributions

Document $m = 42$ (column): Traditional machine learning relies on the availability of a large amount of data to train a model, which is then applied to test data in the same feature space. However, labeled data are often scarce and expensive to obtain...
Strongest topics: $k = \{25, 21, 48, \ldots\}$

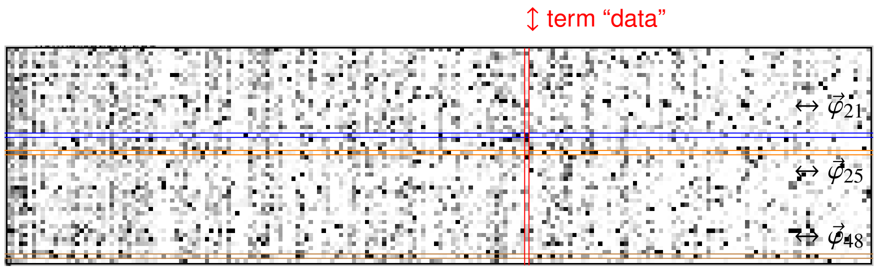


transposed view: rows = topics, columns = documents

Figure: Excerpt from document–topic matrix ϑ $(M = 200, K = 50)$.

Example topic–term distributions

Topic $k = 21$ (row): data word feature label data scarce obtain...
Topic $k = 25$ (row): machine learning train model test feature space...
Topic $k = 48$ (row): computing support grant project system method...



rows = topics, columns = terms

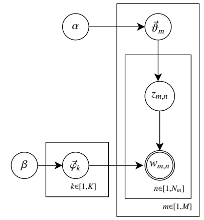
Figure: Excerpt from topic–term matrix φ $(V = 200, K = 50)$.

Example: Text mining for semantic clusters

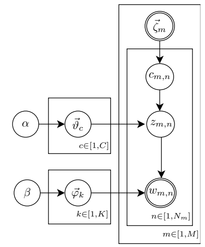
Topic label	Most likely terms according to $\varphi_{k,t} = p(\text{word}\|\text{topic})$
Politische Parteien	CDU Partei Kohl Aufklärung Schäuble Zeitung Union Krise Wahrheit Affäre Christdemokraten Glaubwürdigkeit Konsequenzen
Bundesliga	FC SC München Borussia SV VfL Kickers SpVgg Uhr Köln Bochum Freiburg VfB Eintracht Bayern Hamburger Bayern+München
Polizei / Unfall	Polizei verletzt schwer Auto Unfall Fahrer Angaben schwer+verletzt Menschen Wagen Verletzungen Lawine Mann vier Meter Straße
Tschetschenien	Rebellen russischen Grosny russische Tschetschenien Truppen Kaukasus Moskau Angaben Interfax tschetschenischen Agentur
Politik / Hessen	FDP Koch Hessen CDU Koalition Gerhardt Wagner Liberalen hessischen Westerwelle Wolfgang Roland+Koch Wolfgang+Gerhardt
Wetter	Grad Temperaturen Regen Schnee Süden Norden Sonne Wetter Wolken Deutschland zwischen Nacht Wetterdienst Wind
Politik / Kroatien	Parlament Partei Stimmen Mehrheit Wahlen Wahl Opposition Kroatien Präsident Parlamentswahlen Mesic Abstimmung HDZ
Die Grünen	Grünen Parteitag Atomausstieg Trittin Grüne Partei Trennung Mandat Ausstieg Amt Roestel Jahren Müller Radcke Koalition
Russische Politik	Russland Putin Moskau russischen russische Jelzin Wladimir Tschetschenien Ruslands Wladimir+Putin Kreml Boris Präsidenten
Polizei / Schulen	Polizei Schulen Schüler Täter Polizisten Schule Tat Lehrer erschossen Beamten Mann Polizist Beamte verletzt Waffe

Bigram LDA topics, 18400 German news messages, Jan. 2000 (Heinrich et al. 2005)

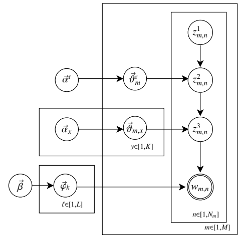
Topic models: Example structures



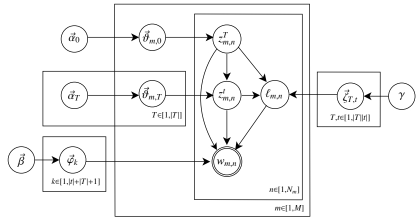
(a) Latent Dirichlet allocation (LDA)



(b) Author–topic model (ATM)



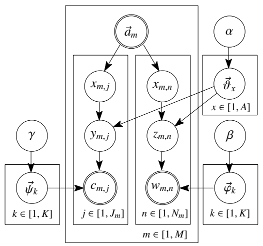
(c) Pachinko allocation model (PAM4)



(d) Hierarchical PAM (hPAM)

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

Typical derivation method (Is it really that complex?)



(e) Expert–tag–topic model (ETT)

(Heinrich 2011)

$$p(\vec{w},\vec{c},\vec{a},\vec{x},\vec{z},\underline{\Theta},\underline{\Phi},\underline{\Psi}|\alpha,\beta,\gamma) = p(\vec{w}|\vec{z},\underline{\Phi})p(\underline{\Phi}|\beta)\cdot p(\vec{c}|\vec{y},\underline{\Psi})p(\underline{\Psi}|\gamma) \cdot p(\vec{y}|\vec{x},\underline{\Theta})p(\vec{z}|\vec{x},\underline{\Theta})p(\underline{\Theta}|\alpha)\cdot p(\vec{x}|\vec{a}) \quad (E.1)$$

$$= \prod_{m=1}^{M}\left(\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}})p(z_{m,n}|\vec{\vartheta}_{x_{m,n}})\,a_{m,x_{m,n}} \cdot \prod_{j=1}^{J_m} p(c_{m,j}|\vec{\psi}_{y_{m,j}})p(y_{m,j}|\vec{\vartheta}_{x_{m,j}})\,a_{m,x_{m,j}}\right) \cdot p(\underline{\Theta}|\alpha)\cdot p(\underline{\Phi}|\beta)\cdot p(\underline{\Psi}|\gamma)\,. \quad (E.2)$$

$$p(\vec{w},\vec{c},\vec{a},\vec{x},\vec{z}|\alpha,\beta,\gamma) = \iiint \prod_{m=1}^{M}\left(\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}})p(z_{m,n}|\vec{\vartheta}_{x_{m,n}})\,a_{m,x_{m,n}} \cdot \prod_{j=1}^{J_m} p(c_{m,j}|\vec{\psi}_{y_{m,j}})p(y_{m,j}|\vec{\vartheta}_{x_{m,j}})\,a_{m,x_{m,j}}\right) \cdot \mathrm{d}p(\underline{\Theta}|\alpha)\cdot \mathrm{d}p(\underline{\Phi}|\beta)\cdot \mathrm{d}p(\underline{\Psi}|\gamma) \quad (E.3)$$

$$= \int \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) \prod_{k=1}^{K} p(\vec{\varphi}_k|\beta)\,\mathrm{d}\varphi_k \cdot \int \prod_{m=1}^{M}\prod_{j=1}^{J_m} p(c_{m,j}|\vec{\psi}_{y_{m,j}}) \prod_{k=1}^{K} p(\vec{\psi}_k|\gamma)\,\mathrm{d}\vec{\psi}_k \cdot \int \prod_{m=1}^{M} p(\underline{\Theta}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_{x_{m,n}})\,a_{m,x_{m,n}} \prod_{j=1}^{J_m} p(y_{m,j}|\vec{\vartheta}_{x_{m,j}})\,a_{m,x_{m,j}}\,\mathrm{d}\vec{\vartheta}_m \quad (E.4)$$

$$= \int \prod_{k=1}^{K} \frac{1}{\Delta_V(\beta)} \prod_{t=1}^{V} \varphi_{k,t}^{n_{k,t}+\beta-1}\,\mathrm{d}\vec{\varphi}_k \cdot \int \prod_{k=1}^{K} \frac{1}{\Delta_C(\gamma)} \prod_{c=1}^{C} \psi_{k,c}^{n_{k,c}+\gamma-1}\,\mathrm{d}\vec{\psi}_k \cdot \int \prod_{a=1}^{A} \frac{1}{\Delta_K(\alpha)} \prod_{k=1}^{K} \vartheta_{a,k}^{n^{(z)}_{a,k}+n^{(y)}_{a,k}+\alpha-1}\,\mathrm{d}\vec{\vartheta}_a \cdot \prod_{m=1}^{M}\prod_{a=1}^{A} a_{m,a}^{n^{(z)}_{m,a}+n^{(y)}_{m,a}} \quad (E.5)$$

$$= \prod_{k=1}^{K} \frac{\Delta(\vec{n}^{(z)}_k+\beta)}{\Delta_V(\beta)} \cdot \frac{\Delta(\vec{n}^{(y)}_k+\gamma)}{\Delta_C(\gamma)} \prod_{a=1}^{A} \frac{\Delta(\vec{n}^{(z)}_a+\vec{n}^{(y)}_a+\alpha)}{\Delta_K(\alpha)} \prod_{m=1}^{M} a_{m,a}^{n^{(z)}_{m,a}+n^{(y)}_{m,a}}\,. \quad (E.6)$$

$$p(z_i{=}k, x_i{=}x|w_i{=}t, \vec{z}_{\neg i}, \vec{y}, \vec{x}_{\neg i}, \vec{w}_{\neg i}, \vec{a}, \vec{c})$$

$$= \frac{p(\vec{w},\vec{z},\vec{y},\vec{x})}{p(\vec{w},\vec{z}_{\neg i},\vec{y},\vec{x}_{\neg i})} = \frac{p(\vec{w}|\vec{z},\vec{y})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i},\vec{y})p(w_i)} \cdot \frac{p(\vec{z}|\vec{x})}{p(\vec{z}_{\neg i}|\vec{x}_{\neg i})} \cdot \frac{p(\vec{x})}{p(\vec{x}_{\neg i})} \quad (E.7)$$

$$\propto \frac{\Delta(\vec{n}^{(z)}_k+\beta)}{\Delta(\vec{n}^{(z)}_{k,\neg i}+\beta)} \cdot \frac{\Delta(\vec{n}_x+\alpha)}{\Delta(\vec{n}_{x,\neg i}+\alpha)} \cdot a_{m,x} \quad (E.8)$$

$$= \frac{\Gamma(n_{k,t}+\beta)\,\Gamma(n_{k,\neg i}+V\beta)}{\Gamma(n_{k,t,\neg i}+\beta)\,\Gamma(n_k+V\beta)} \cdot \frac{\Gamma(n^{(z)}_{x,k}+\alpha)\,\Gamma(n^{(z)}_{x,\neg i}+K\alpha)}{\Gamma(n^{(z)}_{x,k,\neg i}+\alpha)\,\Gamma(n^{(z)}_x+K\alpha)} \cdot a_{m,x} \quad (E.9)$$

$$= \frac{n_{k,t,\neg i}+\beta}{n_{k,\neg i}+V\beta} \cdot \frac{n^{(z)}_{x,k,\neg i}+\alpha}{n^{(z)}_{x,\neg i}+K\alpha} \cdot a_{m,x} \quad (E.10)$$

$$p(y_i{=}k, x_i{=}x|c_i{=}c, \vec{z}_{\neg i}, \vec{y}_{\neg i}, \vec{x}_{\neg i}, \vec{w}, \vec{a}, \vec{c}_{\neg i}) \propto \frac{n_{k,c,\neg i}+\gamma}{n_{k,\neg i}+V\gamma} \cdot \frac{n^{(y)}_{x,k,\neg i}+\alpha}{n^{(y)}_{x,\neg i}+K\alpha} \cdot a_{m,x} \quad (E.12)$$

Topic models – bottom line

- Expanding research field with practical relevance
- No existing analysis as generic model class

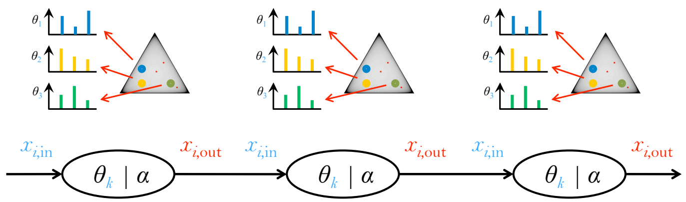
→ Conjecture:

- Important properties generic across models
- Simplifications in the derivation of model properties, inference algorithms and design methods

Overview

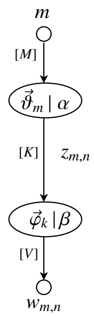
- Topic models – motivation and review
- Networks of mixed membership (NoMMs)
- Inference – a Gibbs "meta-sampler"
- NoMM typology and design
- Application to tag-enhanced expertise finding
- Conclusions and outlook

Generic topic models – "NoMMs"



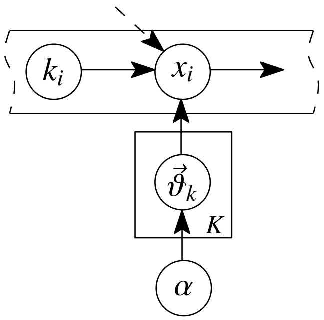
- Generic characteristics of topic models:
  - Levels with multinomial components, generated from Dirichlet
  - Coupling via values of discrete variables

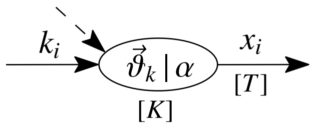
→ "Network of mixed membership" (NoMM): directed graph
  - Compact, domain-specific alternative to Bayesian network
  - Node: Sample from mixture component, selected via incoming edges
  - Terminal node: observation
  - Edge: Propagation of discrete values to children



LDA

NoMM level notation



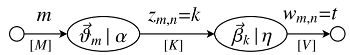


parameters + hyperparameters ⇔ nodes

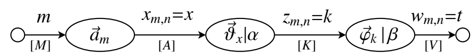
variables ⇔ edges

plates ⇔ indices + dimensions

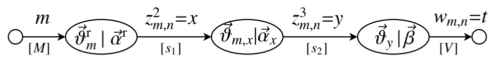
Topic models in NoMM representation



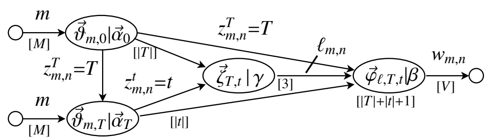
(a) Latent Dirichlet allocation (LDA)



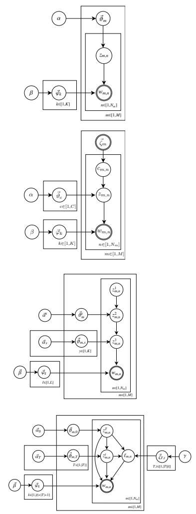
(b) Author–topic model (ATM)



(c) Pachinko allocation model (PAM4)

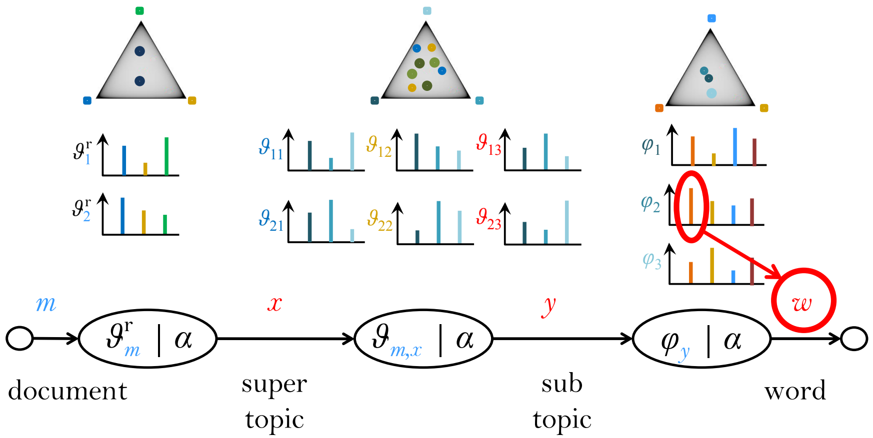


(d) Hierarchical pachinko allocation model (hPAM)



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

Example NoMM generative process: PAM4



Sample from $\varphi_2 \rightarrow$ word $w = 1$.

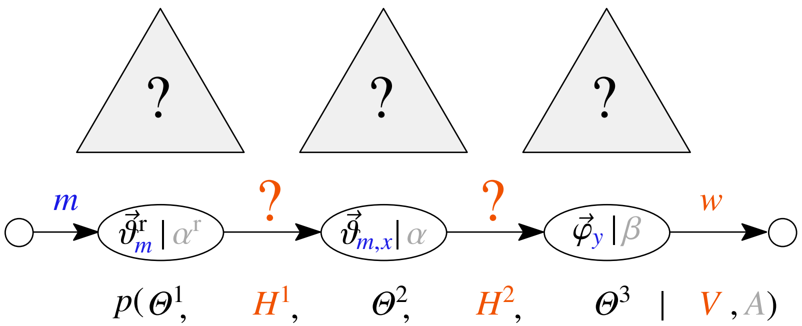
Overview

- Topic models – motivation and review
- Networks of mixed membership (NoMMs)
- Inference – a Gibbs "meta-sampler"
- NoMM typology and design
- Application to tag-enhanced expertise finding
- Conclusions and outlook

Bayesian inference problem

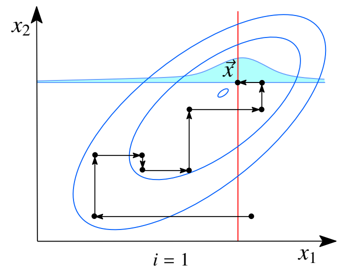
- Bayesian inference: "Reverse generative process"
- Estimate (distributions over) parameters Θ and latent variables ("topics") H given observations V and hyperparameters A.

→ Find posterior distribution $p(H, \Theta \,|\, V, A) \rightarrow$ exponential complexity!



$$p(\Theta^1, \; H^1, \; \Theta^2, \; H^2, \; \Theta^3 \;|\; V, A)$$

Collapsed Gibbs sampling

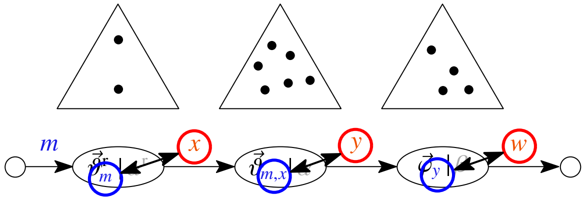


- Collapsed Gibbs sampling: stochastic EM / MCMC:
  - NoMMs: parameters Θ correlated with $H \rightarrow$ integrated out
  - For each data token i: Sample latent variables $H_i = (y_i, z_i, \ldots)$, given all other data, latent $H_{\neg i}$ and visible V:

$$H_i \sim p(H_i \,|\, H_{\neg i}, V, A)\,. \qquad (1)$$

  - Stationary state: full conditional simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational Bayes (Heinrich and Goesele 2009)

Collapsed Gibbs full conditionals



- NoMM full conditionals can be generically derived (Heinrich 2009)
- Typical case leads to weights with straight-forward factor structure:

$$p(H_i \,|\, H_{\neg i}, V, A) \propto \prod_{\ell} \left[\frac{n_{k,t}^{\neg i} + \alpha}{n_k^{\neg i} + T\alpha} \right]^{[\ell]} . \qquad (2)$$

- $n_{k,t}$ = count of co-occurrences between input and output values of a NoMM level ℓ
- More generally: $p(H_i|\cdot) \propto \prod_{\ell} [q(k,t)]^{[\ell]}$ with t = *set* of values/edges

q-functions and Pólya urn

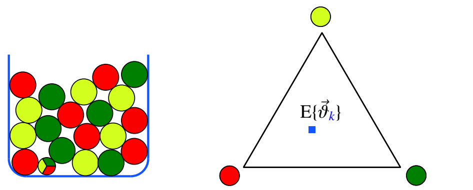


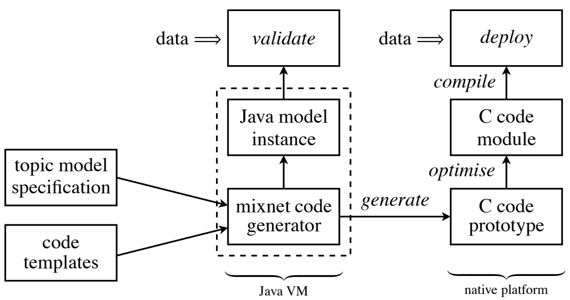
Figure: Pólya urn and multinomial parameters.

$$q(k,t) \triangleq \frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_k^{\neg i} + \alpha)} \overset{|t|=1}{=} \frac{n_{k,t}^{\neg t_i} + \alpha}{n_k^{\neg t_i} + T\alpha} = \text{smoothed ratio of co-occurrence counts}$$

$$\overset{t=\{u,v\}}{=} \frac{n_{k,u}^{\neg u_i} + \alpha}{n_k^{\neg u_i} + T\alpha} \cdot \frac{n_{k,v}^{\neg v_i} + \alpha + \delta(u-v)}{n_k^{\neg v_i} + T\alpha + 1} \triangleq q(k, u \oplus v)$$

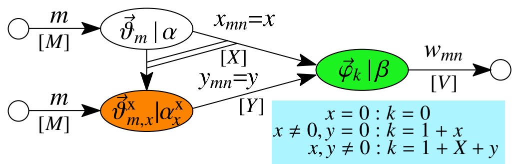
...

Implementation: Gibbs "meta-sampler"



- Code generator for topic models in Java and C
- Separation of knowledge domains: topic model applications vs. machine learning vs. computing architecture

Example NoMM script and generated kernel: hPAM2

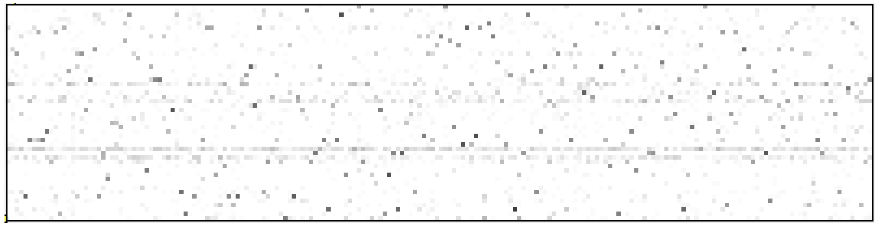



```
model = HPAM2

description:
    Hierarchical PAM model 2 (HPAM2)

sequences:
    # variables sampled for each (m,n)
    w, x, y : m, n

network:
    # each line one NoMM node
    m    >>  theta  | alpha      >>  x
    m,x  >>  thetax | alphax[x]  >>  y
    x,y  >>  phi[k]              >>  w
    # java code to assign k
    k : {
        if (x==0) { k = 0; }
        else if (y==0) k = 1 + x;
        else k = 1 + X + y;
    }.
```

⟶

```
// hidden edge
for (hx = 0; hx < X; hx++) {
    // hidden edge
    for (hy = 0; hy < Y; hy++) {
        mxsel = X * m + hx;
        mxjsel = hx;
        if (hx == 0)
            ksel = 0;
        else if (hy == 0)
            ksel = 1 + hx;
        else
            ksel = 1 + X + hy;
        pp[hx][hy] = (nmx[m][hx] + alpha[hx])
            * (nmxy[mxsel][hy] + alphax[mxjsel][hy])
            / (nmxysum[mxsel] + alphaxsum[mxjsel])
            * (nkw[ksel][w[m][n]] + beta)
            / (nkwsum[ksel] + betasum);
        psum += pp[hx][hy];
    } // for h
} // for h
```

Example document–topic distributions



$t = 500$, converged

Figure: Excerpt from document–topic matrix ϑ ($M = 200, K = 50$).

Fast sampling: hybrid acceleration methods

Serial:

- Exploit saliency of few weights, e.g., generalising (Porteous et al. 2008): compute only few weights on average + estimate normalisation term
- Complex data structures, especially for larger models

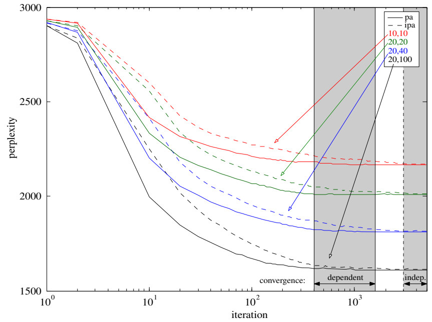
Parallel:

- Distribute local parameters (document-specific etc.)
- Need to sync global parameters: different methods, e.g., generalising (Newman et al. 2009)
- Occupancy: balance communication and computation (architecture-spec.)

Independence assumption:

- Reduce complexity: $\prod_\ell T^\ell \gg \sum_\ell T^\ell$

method	model	parameters	speedup (iter., converge)	
S×P4	LDA	$K = 100$	6.3	
S×P4	LDA	$K = 500$	30.2	
I	PAM4	$K, L = 40, 40$	21.8	7.4
P4×I	PAM4	$K, L = 40, 40$	78.7	24.1
S×P4×I	PAM4	$K, L = 40, 40$	163.2	49.8
S×P4×I	PAM4	$K, L = 20, 100$	143.6	43.5



→ Extend code generation to more complex implementations

Overview

- Topic models – motivation and review
- Networks of mixed membership (NoMMs)
- Inference – a Gibbs “meta-sampler”
- NoMM typology and design
- Application to tag-enhanced expertise finding
- Conclusions and outlook

q-functions and Pólya urn revisited

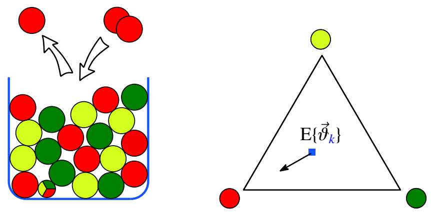


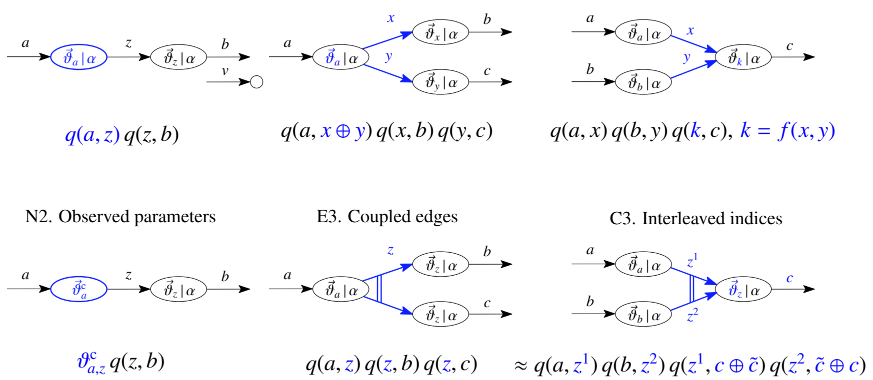
Figure: Pólya urn and multinomial parameters.

$$q(k,t) \triangleq \frac{\mathrm{B}(\vec n_k + \alpha)}{\mathrm{B}(\vec n_k^{\neg i} + \alpha)} \overset{|t|=1}{=} \frac{n_{k,t}^{\neg t_i} + \alpha}{n_k^{\neg t_i} + T\alpha} = \text{smoothed ratio of co-occurrence counts}$$

$$\overset{t=\{u,v\}}{=} \frac{n_{k,u}^{\neg u_i} + \alpha}{n_k^{\neg u_i} + T\alpha} \cdot \frac{n_{k,v}^{\neg v_i} + \alpha + \delta(u-v)}{n_k^{\neg v_i} + T\alpha + 1} \triangleq q(k, u \oplus v)$$

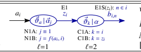
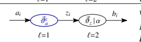
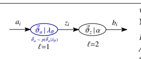
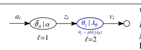
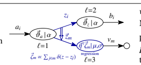
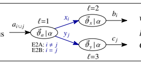
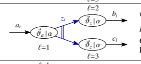
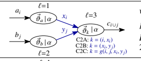
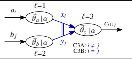
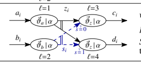
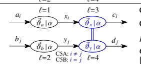
$\ldots$

NoMM sub-structure typology



Gibbs full conditional assembled via:

$$p(H_i \mid \cdot) \propto \prod_{\ell} \left[q(k,t) \right]^{\ell} \quad (3)$$

ID. Name	Structure diagram	Gibbs sampler weight w, Likelihood p for single token i Modelled aspect, example models
N1,E1,C1. Dir–Mult nodes, unbranched		$w(z\vert\cdot) = q(a,z)q(z,b)$ E1S(z_i): $q(a,z)q(z, b_1 \oplus b_2 \oplus \dots b_{N_i})$ $p(b\vert a) = \sum_z \vartheta_{a,z}\vartheta_{z,b}$ *Mixture/admixture*: LDA [Blei et al. 2003b], PAM [Li & McCallum 2006]; LDCC [Shafiei & Milios 2006] (E1S)
N2. Observed parameters		$w(z\vert\vec{\vartheta}^{c}_{a}, \cdot) = \vartheta^{c}_{a,z}\,q(z,b)$ $p(b\vert a) = \sum_z \vartheta^{c}_{a,z}\vartheta_{z,b}$ *Label distribution*: ATM [Rosen-Zvi et al. 2004]
N3. Non-Dirichlet prior		$w(z\vert\vec{\vartheta}_a, \cdot) = p(z_i\vert a_i, \vec{\vartheta}_a)q(z,b)$; M-step: estimate $\vec{\vartheta}_a$ [Blei & Lafferty 2007] $p(b\vert a) = \sum_z \vartheta_{a,z}\vartheta_{z,b}$ *Alternative distributions on the simplex*: CTM [Blei & Lafferty 2007]: $\vec{\vartheta}_a \propto \exp \vec{\eta},\ \vec{\eta} \sim \mathcal{N}(\vec{\mu}, \underline{\Sigma})$; TLM [Wallach 2008]: hierarchy of Dirichlet priors
N4. Non-discrete output		$w(z\vert\theta, \cdot) = q(a,z)p(v_i \mid \theta_z)$; M-step: estimate θ_z $p(v\vert a) = \sum_z \vartheta_{a,z}p(v \mid \theta_z)$ *Non-multinomial observ.*: Corr-LDA [Barnard et al. 2003], GMM [McLachlan & Peel 2000]: $p(v\vert\theta) = \mathcal{N}(\vec{x} \vert \vec{\mu}, \underline{\Sigma})$
N5+E4. Aggregation		$w(z\vert\vec{z}_m, v_m, \cdot) = q(a,z)\,q(z,w)\mathcal{N}(v_m \mid \vec{\eta}_v^{\top}\vec{\zeta}_m, \sigma^2)$; M-step: estimate $\vec{\eta}_v, \sigma^2 \mid \vec{z}, \vec{v}$ (for linear regression, N5B) prediction: $v_m = \vec{\eta}_v^{\top}\vec{\zeta}_m$ *Regression/supervised learning*: Supervised LDA [Blei & McAuliffe 2007], Relational topic model [Chang & Blei 2009]
E2. Autonomous edges		$w(x,y\vert\cdot) = q(a, x\oplus y)q(x,b)q(y,c)$ E2A: $q(a_{i\cup j}, x_i \oplus y_j) = q(a_i, x_i \oplus \tilde{y}_j)q(a_j, \tilde{x}_i \oplus y_j)$ $p(b,c\vert a) = \sum_x \vartheta_{a,x}\vartheta_{x,b} \sum_y \vartheta_{a,y}\vartheta_{y,c}$ *Common mixture of causes*: Multimodal LDA [Ramage et al. 2009]
E3. Coupled edges		$w(z\vert\cdot) = q(a,z)q(z,b)q(z,c)$ $p(b,c\vert a) = \sum_z \vartheta_{a,z}\vartheta_{z,b}\vartheta_{z,c}$ *Common cause for observations*: Hidden relational model (HRM) [Xu et al. 2006], Link-LDA [Erosheva et al. 2004]
C2. Combined indices		$w(x,y\vert\cdot) = q(a,x)q(b,y)q(k,c)$ $p(c\vert a,b) = \sum_x [\vartheta_{a,x} \sum_y \vartheta_{b,y}\vartheta_{k,c}],\ k = g(x,y,i,j)$ *Different dependent causes, relation*: hPAM [Li et al. 2007a], HRM [Xu et al. 2006], Multi-LDA [Porteous et al. 2008a]
C3. Interleaved indices		C3A: $w(x_i, y_j\vert\cdot) = q(a_i, x_i)q(b_j, y_j)q(x_i, c_i \oplus \tilde{c}_j)q(y_j, \tilde{c}_i \oplus c_j)$ C3B: $w(x,y\vert\cdot) = q(a,x)q(b,y)[q(x, c \oplus c)]^{\delta(x-y)}\,[q(x, c \oplus \tilde{c})q(y, \tilde{c} \oplus c)]^{1-\delta(x-y)}$ C3A: $p(c\vert a) = \sum_x \vartheta_{a,x}\vartheta_{x,c}$, $p(c\vert b)$ sim., C3B: $p(c\vert a,b) = \sqrt{\sum_x \vartheta_{a,x}\vartheta_{x,c} \sum_y \vartheta_{b,y}\vartheta_{y,c}}$ *Different causes, same effect*: proposed here
C4. Switch		$w(z,s\vert\cdot) = q(a,z)[q(b,1)q(z,c)]^{\delta(s-1)} \cdot [q(b,2)q(z,d)]^{\delta(s-2)}$ $p(c,d\vert a,b) = \sum_z \vartheta_{a,z}[\vartheta_{b,0}\vartheta_{z,c} + \vartheta_{b,1}\vartheta_{z,d}]$ *Select complex submodels*: Multi-grain LDA [Titov & McDonald 2008], Entity-topic models [Newman et al. 2006a]
C5. Node coupling		C5A: $w(x_i, y_j\vert\cdot) = q(a_i, x_i)q(b_j, y_j)q(x_i, c_i \oplus \tilde{d}_j)q(y_j, \tilde{c}_i \oplus d_j)$ C5B: $w(x,y\vert\cdot) = q(a,x)q(b,y)[q(x, c \oplus d)]^{\delta(x-y)} \cdot [q(x, c \oplus \tilde{d})q(y, \tilde{c} \oplus d)]^{1-\delta(x-y)}$ $p(c,d\vert a,b) = \sum_x \vartheta_{a,x}\vartheta_{x,c} \sum_y \vartheta_{b,y}\vartheta_{y,d}$ *Correlation of submodels, relations*: Simple relational component model [Sinkkonen et al. 2008], Relational topic model [Chang & Blei 2009]

Towards a design process

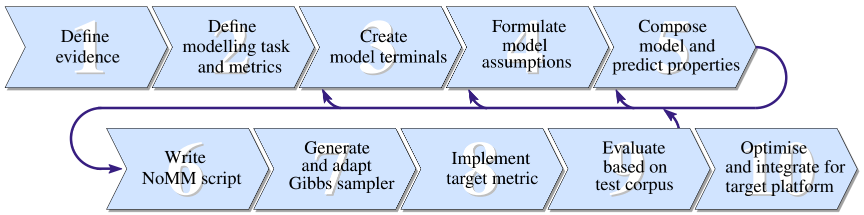


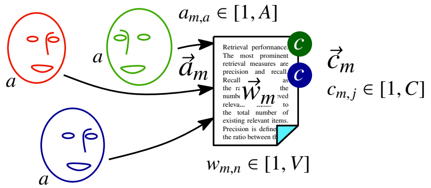
Figure: NoMM design process.

Overview

- Topic models – motivation and review
- Networks of mixed membership (NoMMs)
- Inference – a Gibbs "meta-sampler"
- NoMM typology and design
- Application to tag-enhanced expertise finding
- Conclusions and outlook

Define evidence

- Expertise finding in digital libraries
  - Find authors from document content
  - Semantic tags to disambiguate word meaning and provide additional retrieval method



- Example: scientific community of *Neural Information Processing Systems* (NIPS) conference

Propagation Algorithms for Variational Bayesian Learning

Zoubin Ghahramani and **Matthew J. Beal**
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square, London WC1N 3AR, England
{zoubin,m.beal}@gatsby.ucl.ac.uk

Abstract

Variational approximations are becoming a widespread tool for Bayesian learning of graphical models. We provide some theoretical results for the variational updates in a very general family of *conjugate-exponential* graphical models. We show how the belief propagation and the junction tree algorithms can be used in the inference step of variational Bayesian learning. Applying these results to the Bayesian analysis of linear-Gaussian state-space models we obtain a learning procedure that exploits the Kalman smoothing propagation, while integrating over all model parameters. We demonstrate how this can be used to infer the hidden state dimensionality of the state-space model in a variety of synthetic problems and one real high-dimensional data set.

1 Introduction

Bayesian approaches to machine learning have several desirable properties. Bayesian integration does not suffer overfitting (since nothing is *fit* to the data). Prior knowledge can be incorporated naturally and all uncertainty is manipulated in a consistent manner. Moreover it is possible to learn model structures and readily compare between model classes. Unfortunately, for most models of interest a full Bayesian analysis is computationally intractable.

Until recently, approximate approaches to the intractable Bayesian learning problem had relied either on Markov chain Monte Carlo (MCMC) sampling, the Laplace approximation (Gaussian integration), or asymptotic penalties like BIC. The recent introduction of variational methods for Bayesian learning has resulted in the series of papers showing that these methods can be used to rapidly learn the model structure and approximate the evidence in a wide variety of models. In this paper we will not motivate advantages of the variational Bayesian approach as this is done in

Tags: *probabilistic methods, variational inference, learning algorithms*

Define tasks + metrics; set up terminals

- Retrieval of experts a for term queries $\vec{w}$ and tag queries $\vec{c}$: query likelihood model: $p(\vec{w} \mid a)$ and $p(\vec{c} \mid a)$ → measure retrieval precision
- Topic quality → measure coherence score
- Baseline: Author–topic model ATM (Rosen-Zvi et al. 2004), LDA (Blei et al. 2003)

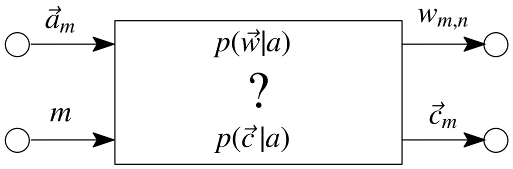
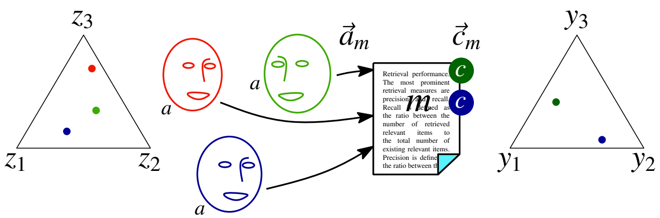


Figure: Model design: Terminals.

Modelling assumptions



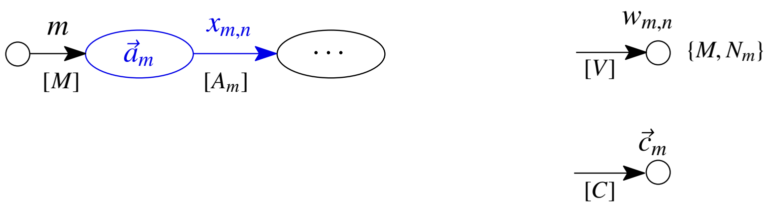
(a) Expertise of authors weighted by the portion of authorship $a_{m,a}$.
(b) Expertise semantics expressed by topics z. Each author has a single field of expertise (topic distribution).
(c) Tag semantics expressed by topics y. Tag topics y could be $\equiv z$.

Compose model

$$p(\ldots \mid \vec{a}, \vec{w}, \vec{c}) \propto \ldots$$

Starting from terminals

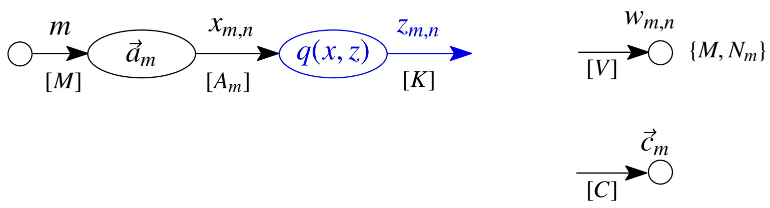
Compose model



$$p(x, \dots \,|\cdot) \propto a_{m,x}\, q(x, \dots) \dots$$

Up-stream evidence $\vec{a}_m$
→ observed parameter node samples word author x

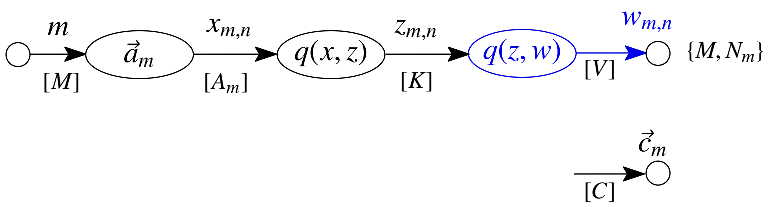
Compose model



$$p(x, z, \dots \,|\cdot) \propto a_{m,x}\, q(x, z) \dots$$

Each author only one field of expertise (topic distribution)
→ q-term $q(x, z)$ assigns topics to sampled author x (cf. ATM)

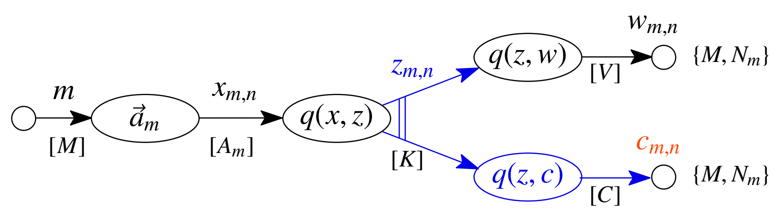
Compose model



$$p(x, z, \dots \,|\cdot) \propto a_{m,x}\, q(x, z)\, q(z, w) \dots$$

Topic distribution over words → can connect directly via $q(z, w)$

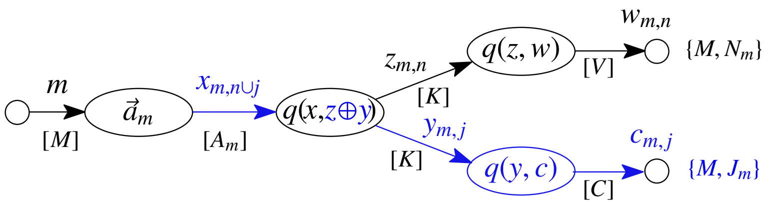
Compose model



$$p(x, z \,|\cdot) \propto a_{m,x}\, q(x, z)\, q(z, w)\, q(z, c)$$

Incorporate tags via $q(z, c)$ conditioned on the same topic
→ Problem: How to determine tag $c_{m,n}$ for word?

Compose model

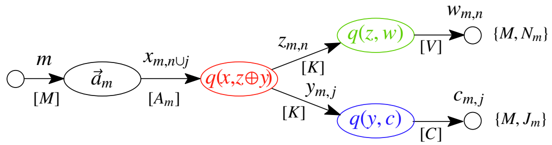


$$p(x,z,y\,|\cdot) \propto a_{m,x}\, q(x,z\oplus y)\, q(z,w)\, q(y,c)$$

→ Incorporate tag topics $y_{m,j}$ on separate sequence (m,j)

→ Tag boosting: adjust tag influence via tag sequence length J_m

ETT1 model



Assembled q-terms:

$$p(x,z,y\,|\cdot) \propto a_{m,x}\, q(x,z\oplus y)\, q(z,w)\, q(y,c) \tag{4}$$

Easy expansion to standard Gibbs full conditionals:

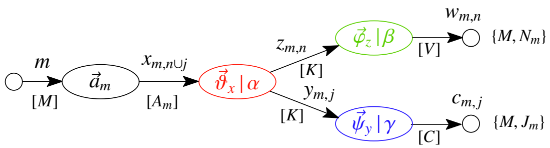
$$p(x_{m,n}{=}x, z_{m,n}{=}z\,|\cdot) \propto a_{m,x} \cdot \frac{n_{x,z}^{\neg\{x,z\}_{m,n}} + \alpha}{n_x^{\neg\{x,z\}_{m,n}} + K\alpha} \cdot \frac{n_{z,w}^{\neg z_{m,n}} + \beta}{n_z^{\neg z_{m,n}} + V\beta} \tag{5}$$

$$p(x_{m,j}{=}x, y_{m,j}{=}y\,|\cdot) \propto a_{m,x} \cdot \frac{n_{x,y}^{\neg\{x,y\}_{m,j}} + \alpha}{n_y^{\neg\{x,y\}_{m,j}} + K\alpha} \cdot \frac{n_{y,c}^{\neg y_{m,j}} + \gamma}{n_y^{\neg y_{m,j}} + C\gamma} \tag{6}$$

Retrieval via query likelihood model:

$$p(\vec{w}\,|\,a) = \textstyle\prod_{w\in\vec{w}} \sum_z \vartheta_{a,z}\varphi_{z,w} \qquad p(\vec{c}\,|\,a) = \prod_{c\in\vec{c}} \sum_y \vartheta_{a,y}\psi_{y,c}\,. \tag{7}$$

ETT1 model



Assembled q-terms:

$$p(x,z,y\,|\cdot) \propto a_{m,x}\, q(x,z\oplus y)\, q(z,w)\, q(y,c) \tag{4}$$

Easy expansion to standard Gibbs full conditionals:

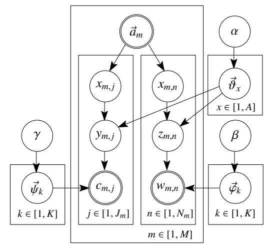
$$p(x_{m,n}{=}x, z_{m,n}{=}z\,|\cdot) \propto a_{m,x} \cdot \frac{n_{x,z}^{\neg\{x,z\}_{m,n}} + \alpha}{n_x^{\neg\{x,z\}_{m,n}} + K\alpha} \cdot \frac{n_{z,w}^{\neg z_{m,n}} + \beta}{n_z^{\neg z_{m,n}} + V\beta} \tag{5}$$

$$p(x_{m,j}{=}x, y_{m,j}{=}y\,|\cdot) \propto a_{m,x} \cdot \frac{n_{x,y}^{\neg\{x,y\}_{m,j}} + \alpha}{n_y^{\neg\{x,y\}_{m,j}} + K\alpha} \cdot \frac{n_{y,c}^{\neg y_{m,j}} + \gamma}{n_y^{\neg y_{m,j}} + C\gamma} \tag{6}$$

Retrieval via query likelihood model:

$$p(\vec{w}\,|\,a) = \textstyle\prod_{w\in\vec{w}} \sum_z \vartheta_{a,z}\varphi_{z,w} \qquad p(\vec{c}\,|\,a) = \prod_{c\in\vec{c}} \sum_y \vartheta_{a,y}\psi_{y,c}\,. \tag{7}$$

Typical derivation method (Is it really that complex?)



(a) Expert–tag–topic model 1 (ETT1)

(Heinrich 2011)

$$p(\vec{w},\vec{c},\vec{a},\vec{x},\vec{z},\underline{\Theta},\underline{\Phi},\underline{\Psi}|\alpha,\beta,\gamma) = p(\vec{w}|\vec{z},\underline{\Phi})p(\underline{\Phi}|\beta)\cdot p(\vec{c}|\vec{y},\underline{\Psi})p(\underline{\Psi}|\gamma) \cdot p(\vec{y}\,|\vec{x},\underline{\Theta})p(\vec{z}\,|\vec{x},\underline{\Theta})p(\underline{\Theta}|\alpha)\cdot p(\vec{x}\,|\vec{a}) \tag{E.1}$$

$$= \prod_{m=1}^{M}\Big(\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}})p(z_{m,n}|\vec{\vartheta}_{x_{m,n}})\,a_{m,x_{m,n}} \cdot \prod_{j=1}^{J_m} p(c_{m,j}|\vec{\psi}_{y_{m,j}})p(y_{m,j}|\vec{\vartheta}_{x_{m,j}})\,a_{m,x_{m,j}}\Big) \cdot p(\underline{\Theta}|\alpha)\cdot p(\underline{\Phi}|\beta)\cdot p(\underline{\Psi}|\gamma)\,. \tag{E.2}$$

$$p(\vec{w},\vec{c},\vec{a},\vec{x},\vec{z}|\alpha,\beta,\gamma) = \iiint \prod_{m=1}^{M}\Big(\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}})p(z_{m,n}|\vec{\vartheta}_{x_{m,n}})\,a_{m,x_{m,n}} \cdot \prod_{j=1}^{J_m} p(c_{m,j}|\vec{\psi}_{y_{m,j}})p(y_{m,j}|\vec{\vartheta}_{x_{m,j}})\,a_{m,x_{m,j}}\Big) \cdot \mathrm{d}p(\underline{\Theta}|\alpha)\cdot \mathrm{d}p(\underline{\Phi}|\beta)\cdot \mathrm{d}p(\underline{\Psi}|\gamma) \tag{E.3}$$

$$= \int \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) \prod_{k=1}^{K} p(\vec{\varphi}_k|\beta)\,\mathrm{d}\varphi_k \cdot \int \prod_{m=1}^{M}\prod_{j=1}^{J_m} p(c_{m,j}|\vec{\psi}_{y_{m,j}}) \prod_{k=1}^{K} p(\vec{\psi}_k|\gamma)\,\mathrm{d}\vec{\psi}_k \cdot \int \prod_{m=1}^{M} p(\underline{\Theta}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_{x_{m,n}})\,a_{m,x_{m,n}} \prod_{j=1}^{J_m} p(y_{m,j}|\vec{\vartheta}_{x_{m,j}})\,a_{m,x_{m,j}}\,\mathrm{d}\vec{\vartheta}_m \tag{E.4}$$

$$= \int \prod_{k=1}^{K} \frac{1}{\Delta_V(\beta)} \prod_{t=1}^{V} \varphi_{k,t}^{n_{k,t}+\beta-1}\,\mathrm{d}\vec{\varphi}_k \cdot \int \prod_{k=1}^{K} \frac{1}{\Delta_C(\gamma)} \prod_{c=1}^{C} \psi_{k,c}^{n_{k,c}+\gamma-1}\,\mathrm{d}\vec{\psi}_k \cdot \int \prod_{a=1}^{A} \frac{1}{\Delta_K(\alpha)} \prod_{k=1}^{K} \vartheta_{a,k}^{n_{a,k}^{(z)}+n_{a,k}^{(y)}+\alpha-1}\,\mathrm{d}\vec{\vartheta}_a \cdot \prod_{m=1}^{M}\prod_{a=1}^{A} a_{m,a}^{n_{m,a}^{(z)}+n_{m,a}^{(y)}} \tag{E.5}$$

$$= \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k^{(z)}+\beta)}{\Delta_V(\beta)} \cdot \frac{\Delta(\vec{n}_k^{(y)}+\gamma)}{\Delta_C(\gamma)} \prod_{a=1}^{A} \frac{\Delta(\vec{n}_a^{(z)}+\vec{n}_a^{(y)}+\alpha)}{\Delta_K(\alpha)} \prod_{m=1}^{M} a_{m,a}^{n_{m,a}^{(z)}+n_{m,a}^{(y)}}\,. \tag{E.6}$$

$$p(z_i{=}k, x_i{=}x|w_i{=}t, \vec{z}_{\neg i}, \vec{y}, \vec{x}_{\neg i}, \vec{w}_{\neg i}, \vec{a}, \vec{c}) = \frac{p(\vec{w},\vec{z},\vec{y},\vec{x})}{p(\vec{w},\vec{z}_{\neg i},\vec{y},\vec{x}_{\neg i})} = \frac{p(\vec{w}|\vec{z},\vec{y})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i},\vec{y})p(w_i)} \cdot \frac{p(\vec{z}\,|\vec{x})}{p(\vec{z}_{\neg i}|\vec{x}_{\neg i})} \cdot \frac{p(\vec{x})}{p(\vec{x}_{\neg i})} \tag{E.7}$$

$$\propto \frac{\Delta(\vec{n}_k^{(z)}+\beta)}{\Delta(\vec{n}_{k,\neg i}^{(z)}+\beta)} \cdot \frac{\Delta(\vec{n}_x+\alpha)}{\Delta(\vec{n}_{x,\neg i}+\alpha)} \cdot a_{m,x} \tag{E.8}$$

$$= \frac{\Gamma(n_{k,t}+\beta)\,\Gamma(n_{k,\neg i}+V\beta)}{\Gamma(n_{k,t,\neg i}+\beta)\,\Gamma(n_k+V\beta)} \cdot \frac{\Gamma(n_{x,k}^{(z)}+\alpha)\,\Gamma(n_{x,\neg i}^{(z)}+K\alpha)}{\Gamma(n_{x,k,\neg i}^{(z)}+\alpha)\,\Gamma(n_x^{(z)}+K\alpha)} \cdot a_{m,x} \tag{E.9}$$

$$= \frac{n_{k,t,\neg i}+\beta}{n_{k,\neg i}+V\beta} \cdot \frac{n_{x,k,\neg i}^{(z)}+\alpha}{n_{x,\neg i}^{(z)}+K\alpha} \cdot a_{m,x} \tag{E.10}$$

$$p(y_i{=}k, x_i{=}x|c_i{=}c, \vec{z}_{\neg i}, \vec{y}_{\neg i}, \vec{x}_{\neg i}, \vec{w}, \vec{a}, \vec{c}_{\neg i}) \propto \frac{n_{k,c,\neg i}+\gamma}{n_{k,\neg i}+V\gamma} \cdot \frac{n_{x,k,\neg i}^{(y)}+\alpha}{n_{x,\neg i}^{(y)}+K\alpha} \cdot a_{m,x} \tag{E.12}$$

Model evaluation

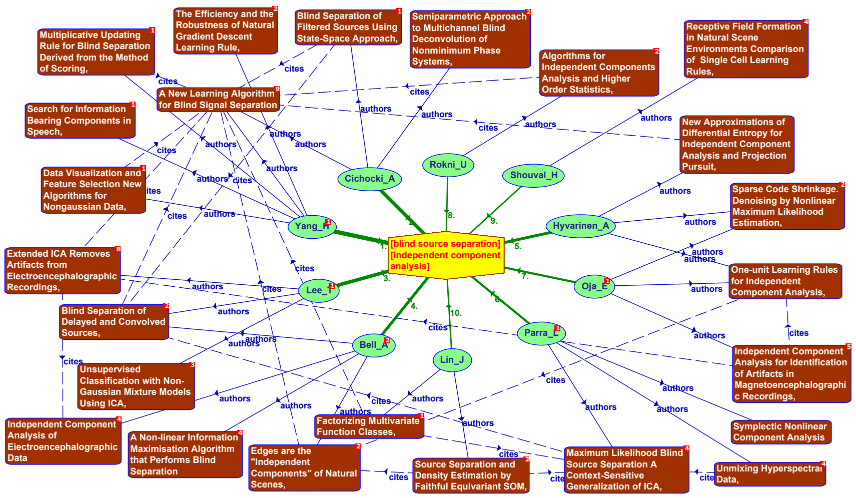
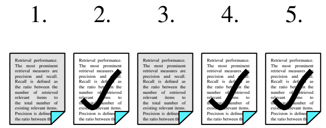
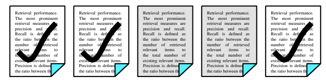


Figure: ETT1 example query in community browser.

Truncated average precision



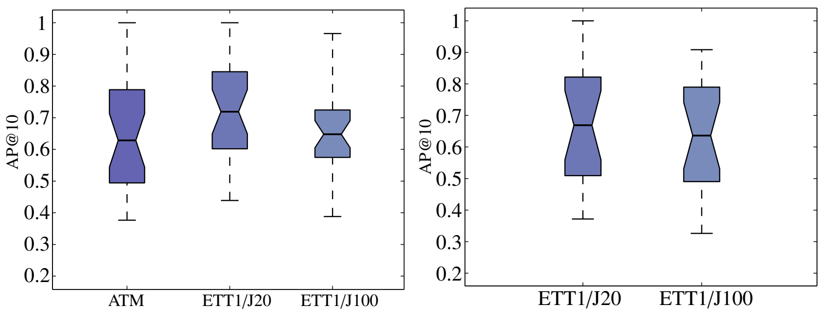
$$AP@5 = \frac{1/2 \quad +2/4 \quad +3/5}{3} = 0.533$$



$$AP@5 = \frac{1/1 \quad +2/2 \quad +3/5}{3} = 0.867$$

Figure: Average precision at 5 (3 relevant documents in corpus).

Retrieval and clustering results



(a) Term queries (b) Tag queries

- Term retrieval improved by tag influence during *training* time
- Mutual information between a-priori tag clusterings $p(c|a)$ and topic clusterings $p(z|a)$: ETT1 ≥ 1.002 vs. ATM $= 0.865$.
- Semi-supervised features: find relevant items with missing tags
- Tag strength: bias towards strong tags in combinations

Topic coherence results

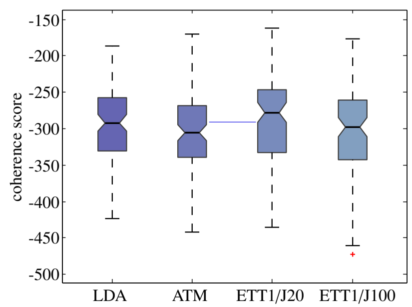
Topic coherence (Mimno et al. 2011):

- ≈ How often do top-ranked topic terms co-occur in documents?
- Re-enacts human judgement in topic intrusion experiments (Chang et al. 2009; Heinrich 2011)

Words in topic (choose worst match (A-F) in every group):

1. A. orientation	2. A. likelihood	3. A. risk
B. cortex	B. mixture	B. return
C. visual	C. theorem	C. stock
D. ocular	D. density	D. trading
E. acoustic	E. em	E. processor
F. eye	F. prior	F. prediction
4. A. language	5. A. circuit	6. A. validation
B. word	B. bayesian	B. set
C. stress	C. analog	C. variance
D. grammar	D. voltage	D. regression
E. neural	E. vlsi	E. selection
F. syllable	F. chip	F. bias

(a) Topic intrusion experiment



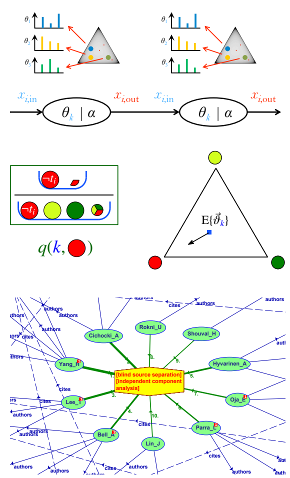
(b) Coherence scores

Overview

- Topic models – motivation and review
- Networks of mixed membership (NoMMs)
- Inference – a Gibbs “meta-sampler”
- NoMM typology and design
- Application to tag-enhanced expertise finding
- Conclusions and outlook

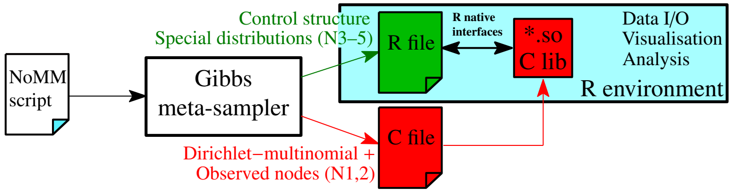
Conclusions

- Networks of mixed membership: Domain-specific compact representation
- Inference:
  - Generic Gibbs sampling: q-functions as central quantity in model behaviour
  - Gibbs meta-sampler: simplify implementation
  - Hybrid acceleration methods
  - Alternatives: variational Bayes (Heinrich and Goesele 2009), collapsed VB
- Typology and design method:
  - Model structure types: literature + novel
  - Building blocks for design with predictable properties
- Application:
  - Expert–tag–topic model demonstrates design
  - Tags improve retrieval and topic coherence



Towards an R-based Gibbs meta-sampler

- R environment becoming popular for topic models, e.g.:
  - `topicmodels` package implementing general and various special cases (Grün and Hornik 2011), based on text mining package `tm`
  - `lda` package with LDA, supervised, relational topic models (Blei et al. 2003; Blei and McAuliffe 2007; Chang and Blei 2009)
- Vision: Use Gibbs meta-sampler as front-end to create R-based high-performance code ↔ use R as experimental front-end



- Extend to non-parametric distributions, e.g., based on `DPpackage` (Jara et al. 2012):
  - NoMMs as polymorphism of parametric and non-parametric models (with different Bayesian networks)

Q+A

http://arbylon.net/resources.html

References I

References

Blei, D. and J. McAuliffe (2007).
Supervised topic models.
In *Advances in Neural Information Processing Systems.*

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

Buntine, W. and A. Jakulin (2005).
Discrete principal components analysis.
In *Proc. ECML.*

Chang, J. and D. M. Blei (2009).
Relational topic models for document networks.
In *AISTATS.*

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei (2009).
Reading tea leaves: How humans interpret topic models.
In *Proc. Neural Information Processing Systems (NIPS).*

Griffiths, T. L., J. B. Tenenbaum, and M. Steyvers (2007).
Topics in semantic representation.
Psychological Review 114(2), 211–244.

References II

Grün, B. and K. Hornik (2011).
topicmodels: An R package for fitting topic models.
Journal of Statistical Software 43(13).

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), Part 1*, pp. 517–532.

Heinrich, G. (2011).
Typology of mixed-membership models: Towards a design method.
In *Proc. European Conf. on Mach. Learn. / Principles and Pract. of Know. Discov. in Databases (ECML/PKDD).*

Heinrich, G. and M. Goesele (2009).
Variational Bayes for generic topic models.
In *Proc. 32nd Annual German Conference on Artificial Intelligence (KI2009).*

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.*

References III

Jara, A., T. Hanson, F. A. Quintana, P. Mueller, and G. L. Rosner (2012, Feb.).
DPpackage: Bayesian nonparametric modeling in R.
software documentation.

Landauer, T. K. and S. T. Dumais (1997).
Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.
Psych. Rev. 104(2), 211–240.
Cognitive view on LSA.

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

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.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011, July).
Optimizing semantic coherence in topic models.
In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK*, pp. 262272.

References IV

Newman, D., A. Asuncion, P. Smyth, and M. Welling (2009, August).
Distributed algorithms for topic models.
JMLR 10, 1801–1828.

Porteous, I., D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling (2008).
Fast collapsed Gibbs sampling for latent Dirichlet allocation.
In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 569–577. ACM.

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