A generic approach to topic models and its application to virtual communities

Gregor Heinrich

PhD presentation (English translation incl. backup slides, 45min)
Faculty of Mathematics and Computer Science
University of Leipzig

28 November 2012

Version 2.9 EN BU
A generic approach to topic models and its application to virtual communities

Gregor Heinrich

PhD presentation (English translation incl. backup slides, 45min)
Faculty of Mathematics and Computer Science
University of Leipzig

28 November 2012

Version 2.9 EN BU
Overview

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook
Motivation: Virtual communities

- Virtual communities = groups of persons who exchange information and knowledge electronically
- Examples: organisations, digital libraries, “Web 2.0” applications incl. social networks
- Data are multimodal: text content; authorship, citation, annotations and recommendations; cooperation and other social relations
- Typical case: discrete data with high dynamics and large volumes
Motivation: Unsupervised mining of discrete data

- Identification of relationships in large data volumes
- Only data (and possibly model) required (information retrieval, network analysis, clustering, NLP methods)
- **Density problem**: Features too sparse for analysis in high-dimensional feature space
- **Vocabulary problem**: Semantic similarity ≠ lexical similarity (polysemy, synonymy, etc.)
Motivation: Unsupervised mining of discrete data

- Identification of relationships in large data volumes
- Only data (and possibly model) required (information retrieval, network analysis, clustering, NLP methods)
- Density problem: Features too sparse for analysis in high-dimensional feature space
- Vocabulary problem: Semantic similarity $\neq$ lexical similarity (polysemy, synonymy, etc.)
Topic models as approach

- Probabilistic representations of grouped discrete data
- Illustrative for text: Words grouped in documents
  - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  - Language = Mixture of topics (latent semantic structure)

→ Reduce vocabulary problem: Find semantic relations
→ Reduce density problem: Dimensionality reduction
Probabilistic representations of grouped discrete data

Illustrative for text: Words grouped in documents
  - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  - Language = Mixture of topics (latent semantic structure)

→ Reduce vocabulary problem: Find semantic relations
→ Reduce density problem: Dimensionality reduction
Probabilistic representations of grouped discrete data
Illustrative for text: Words grouped in documents
- Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
- Language = Mixture of topics (latent semantic structure)

→ Reduce vocabulary problem: Find semantic relations
→ Reduce density problem: Dimensionality reduction
**Topic models as approach**

- Probabilistic representations of grouped discrete data
- Illustrative for text: Words grouped in documents
  - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  - Language = Mixture of topics (latent semantic structure)
- Reduce vocabulary problem: Find semantic relations
- Reduce density problem: Dimensionality reduction
Topic models as approach

- Probabilistic representations of grouped discrete data
- Illustrative for text: Words grouped in documents
  - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  - Language = Mixture of topics (latent semantic structure)

→ Reduce vocabulary problem: Find semantic relations
→ Reduce density problem: Dimensionality reduction
Probabilistic representations of grouped discrete data
Illustrative for text: Words grouped in documents
  Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  Language = Mixture of topics (latent semantic structure)

→ Reduce **vocabulary problem**: Find semantic relations
→ Reduce **density problem**: Dimensionality reduction
Topic models as approach

- Probabilistic representations of grouped discrete data
- Illustrative for text: Words grouped in documents
  - Latent Topics = Probability distributions over vocabulary. Dominant terms of a topic are semantically similar.
  - Language = Mixture of topics (latent semantic structure)
- Reduce vocabulary problem: Find semantic relations
- Reduce density problem: Dimensionality reduction
Language models: Unigram model

One distribution for all data
Language models: Unigram mixture model

$p(w | z_1)$

$p(w | z_2)$

$z_1$

$z_2$

$w_{1,1}$ $w_{1,2}$ $w_{1,3}$

document 1

$w_{2,1}$ $w_{2,2}$ $w_{2,3}$

document 2

$z_m$

$W_{m,n}$

word $n$

document $m$

One distribution per document
Language models: Unigram admixture model

One distribution per word → basic topic model
Language models: Unigram admixture model

One distribution per word → basic topic model
Bayesian topic models: The Dirichlet distribution

Bayesian methodology:
- Distributions generated from prior distributions
- Speech + other discrete data: Dirichlet distribution

Important prior:
- Defined on simplex: Surface containing all discrete distributions
- Parameter $\vec{\alpha}$ controls behaviour

$\vec{\alpha} = (4, 4, 2)$

Bayesian topic model: Latent Dirichlet Allocation (LDA) (Blei et al. 2003)
Latent Dirichlet Allocation (Blei et al. 2003)
Latent Dirichlet Allocation (Blei et al. 2003)

Concert tonight at Rhythm and Spice Restaurant...
Latent Dirichlet Allocation

Concert tonight at Rhythm and Spice Restaurant...

Generating word distributions for all topics
Generating topic distribution for document

Concert tonight at Rhythm and Spice Restaurant ...
Latent Dirichlet Allocation

Concert tonight at Rhythm and Spice Restaurant . . .

Sampling the topic index for first word, \( z = 2 \)
Sampling a word from term distribution for topic 2, “concert”
State of the art

Large number of published models that extend LDA:
- Authors (Rosen-Zvi et al. 2004),
- Citations (Dietz et al. 2007),
- Hierarchy (Li and McCallum 2006; Li et al. 2007),
- Image features and captions (Barnard et al. 2003) etc.
- Results for “topic model” (title + abstract) only since 2012: ACM >400, Google Scholar >1300.

→ Expanding research area with practical relevance
- But: No existing analysis as generic model class
- Partly tedious derivation, especially for inference algorithms

Conjecture:
- Important properties generic across models
- Simplifications for derivation of concrete model properties, inference algorithms and design methods
Large number of published models that extend LDA:

Authors (Rosen-Zvi et al. 2004), Citations (Dietz et al. 2007), Hierarchy (Li and McCallum 2006; Li et al. 2007), Image features and captions (Barnard et al. 2003) etc.

Results for "topic model" (title + abstract) only since 2012: ACM > 400, Google Scholar > 1300.

Expanding research area with practical relevance but: No existing analysis as generic model class

Partly tedious derivation, especially for inference algorithms

Conjecture:

Important properties generic across models

Simplifications for derivation of concrete model properties, inference algorithms and design methods

---

Expert–tag–topic model

(Heinrich 2011)

\[
p(\vec{w}, \vec{d}, \vec{X}, \vec{Y}, \beta, \gamma) = \prod_{m=1}^{M} \left( \prod_{n=1}^{N_m} p(w_{m,n} | \vec{\phi}_m) \prod_{j=1}^{J_m} p(y_{m,j} | \vec{\psi}_m) \right) \prod_{k=1}^{K} p(c_k | \vec{\varphi}_k)
\]

(E.1)

\[
p(\vec{d}, \vec{X}, \vec{Y}, \beta, \gamma)
\]

(E.2)

\[
p(\vec{d}, \vec{X}, \vec{Y}) = \prod_{k=1}^{K} \prod_{j=1}^{J_m} \frac{\Gamma(n_{k,j} + 1)}{\Gamma(n_{k,j} + 1) \Gamma(\alpha_{k,j} + 1) \cdot \Delta_{\alpha_{k,j}}(\alpha_{k,j})}
\]

(E.8)

\[
p(\vec{w}, \vec{d}, \vec{X}, \vec{Y}, \vec{Z}, \vec{\varphi}, \vec{\psi}, \vec{\alpha}, \vec{\beta}, \vec{\gamma})
\]

(E.9)

\[
p(\vec{w}, \vec{d}, \vec{X}, \vec{Y}, \vec{Z}, \vec{\varphi}, \vec{\psi}, \vec{\alpha}, \vec{\beta}, \vec{\gamma})
\]

(E.10)

\[
p(\vec{w}, \vec{d}, \vec{X}, \vec{Y}, \vec{Z}, \vec{\varphi}, \vec{\psi}, \vec{\alpha}, \vec{\beta}, \vec{\gamma})
\]

(E.11)

\[
p(\vec{w}, \vec{d}, \vec{X}, \vec{Y}, \vec{Z}, \vec{\varphi}, \vec{\psi}, \vec{\alpha}, \vec{\beta}, \vec{\gamma})
\]

(E.12)
State of the art

Large number of published models that extend LDA:
- Authors (Rosen-Zvi et al. 2004),
- Citations (Dietz et al. 2007),
- Hierarchy (Li and McCallum 2006; Li et al. 2007),
- Image features and captions (Barnard et al. 2003) etc.
- Results for “topic model” (title + abstract) only since 2012: ACM >400, Google Scholar >1300.

→ Expanding research area with practical relevance
  - But: No existing analysis as generic model class
  - Partly tedious derivation, especially for inference algorithms

Conjecture:
- Important properties generic across models
- Simplifications for derivation of concrete model properties, inference algorithms and design methods
State of the art

Large number of published models that extend LDA:
- Authors (Rosen-Zvi et al. 2004),
- Citations (Dietz et al. 2007),
- Hierarchy (Li and McCallum 2006; Li et al. 2007),
- Image features and captions (Barnard et al. 2003) etc.
- Results for “topic model” (title + abstract) only since 2012: ACM >400, Google Scholar >1300.

→ Expanding research area with practical relevance
  - But: No existing analysis as generic model class
  - Partly tedious derivation, especially for inference algorithms

Conjecture:
- Important properties generic across models
- Simplifications for derivation of concrete model properties, inference algorithms and design methods
Research questions

“How can topic models be described in a generic way in order to use their properties across particular applications?”

“Can generic topic models be implemented generically and, if so, can repeated structures be exploited for optimisations?”

“How can generic models be applied to data in virtual communities?”
Overview

- Introduction
- **Generic topic models**
- Inference methods
- Application to virtual communities
- Conclusions and outlook
“How can topic models be described in a generic way in order to use their properties across particular applications?”
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)
Generic topic models – “NoMMs”

- Generic characteristics of topic models:
  - Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
  - Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclic graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.
Generic characteristics of topic models:

- Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
- Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):

- Compact representation for topic models
- Directed acyclical graph
- Node: sample from mixture component, selection via incoming edges; terminal node: observation
- Edge: propagation of discrete values to child nodes.
Generic characteristics of topic models:

- Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
- Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):

- Compact representation for topic models
- Directed acyclic graph
- Node: sample from mixture component, selection via incoming edges; terminal node: observation
- Edge: propagation of discrete values to child nodes.
Generic topic models – “NoMMs”

- Generic characteristics of topic models:
  - Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
  - Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):

  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.
Generic topic models – “NoMMs”

- Generic characteristics of topic models:
  - Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
  - Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.
Generic topic models – “NoMMs”

- **Generic characteristics of topic models:**
  - Levels with discrete components $\vec{\theta}_k$, generated from Dirichlet distributions
  - Coupling via values of discrete variables $x$

- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.
Generic topic models – “NoMMs”

- **Generic characteristics of topic models:**
  - Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
  - Coupling via values of discrete variables $x$

- “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclical graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.
Generic topic models – “NoMMs”

- Generic characteristics of topic models:
  - Levels with discrete components \( \tilde{\theta}_k \), generated from Dirichlet distributions
  - Coupling via values of discrete variables \( x \)

→ “Network of mixed membership” (NoMM):
  - Compact representation for topic models
  - Directed acyclic graph
  - Node: sample from mixture component, selection via incoming edges; terminal node: observation
  - Edge: propagation of discrete values to child nodes.
Generic characteristics of topic models:

- Levels with discrete components $\tilde{\theta}_k$, generated from Dirichlet distributions
- Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):

- Compact representation for topic models
- Directed acyclical graph
- Node: sample from mixture component, selection via incoming edges; terminal node: observation
- Edge: propagation of discrete values to child nodes.
Generic characteristics of topic models:

- Levels with discrete components $\bar{\theta}_k$, generated from Dirichlet distributions
- Coupling via values of discrete variables $x$

→ “Network of mixed membership” (NoMM):

- Compact representation for topic models
- Directed acyclical graph
- Node: sample from mixture component, selection via incoming edges; terminal node: observation
- Edge: propagation of discrete values to child nodes.
Generic topic models – “NoMMs”

Generic characteristics of topic models:
- Levels with discrete components $\bar{\theta}_k$, generated from Dirichlet distributions
- Coupling via values of discrete variables $\chi$

→ “Network of mixed membership” (NoMM):
- Compact representation for topic models
- Directed acyclical graph
- Node: sample from mixture component, selection via incoming edges; terminal node: observation
- Edge: propagation of discrete values to child nodes.
Topic models as NoMMs

(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

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook
“Can generic topic models be implemented generically...?”
Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters $\Theta$ and latent variables/topics $H$, given observations $V$ and Dirichlet parameters $A$
  - Determine posterior distribution $p(H, \Theta | V, A)$
- Intractability $\rightarrow$ approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters $\Theta$ (“Collapsed” GS)
  - Sample topics $H_i$ for each data point $i$ in turn: $H_i \sim p(H_i | H_{-i}, V, A)$
Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters $\Theta$ and latent variables/topics $H$, given observations $V$ and Dirichlet parameters $A$
  - Determine posterior distribution $p(H, \Theta | V, A)$
- Intractability $\rightarrow$ approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters $\Theta$ (“Collapsed” GS)
  - Sample topics $H_i$ for each data point $i$ in turn: $H_i \sim p(H_i | H_{-i}, V, A)$
Bayesian inference problem and Gibbs sampler

- Bayesian inference: "inversion of generative process":
  - Find distributions over parameters $\Theta$ and latent variables/topics $H$, given observations $V$ and Dirichlet parameters $A$
  - Determine posterior distribution $p(H, \Theta | V, A)$

- Intractability $\rightarrow$ approximative approaches

- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters $\Theta$ ("Collapsed" GS)
  - Sample topics $H_i$ for each data point $i$ in turn: $H_i \sim p(H_i | H_{-i}, V, A)$

Gregor Heinrich
A generic approach to topic models
Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters $\Theta$ and latent variables/topics $H$, given observations $V$ and Dirichlet parameters $A$
  - Determine posterior distribution $p(H, \Theta | V, A)$

- Intractability $\rightarrow$ approximative approaches
  - Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
    - In topic models: Marginalise parameters $\Theta$ (“Collapsed” GS)
    - Sample topics $H_i$ for each data point $i$ in turn: $H_i \sim p(H_i | H_{\neg i}, V, A)$

\[
p(\Theta^1, H^1, \Theta^2, H^2, \Theta^3 | V, A)
\]
Bayesian inference problem and Gibbs sampler

- Bayesian inference: “inversion of generative process”:
  - Find distributions over parameters $\Theta$ and latent variables/topics $H$, given observations $V$ and Dirichlet parameters $A$
  - Determine posterior distribution $p(H, \Theta | V, A)$

- Intractability $\rightarrow$ approximative approaches
- Gibbs sampling: Variant of Markov-Chain Monte Carlo (MCMC)
  - In topic models: Marginalise parameters $\Theta$ (“Collapsed” GS)
  - Sample topics $H_i$ for each data point $i$ in turn: $H_i \sim p(H_i | H_{\neg i}, V, A)$
Gibbs sampler can be generically derived (Heinrich 2009)

Typical case: Quotients of factors over levels $\ell$:

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

- $n_{k,t} = \text{count of co-occurrences between input and output values of a level (components and samples)}$

- More complex variants covered by $q(k, t) \triangleq \frac{\text{beta}(\{n_{k,t}\}_t^{T} + \alpha)}{\text{beta}(\{n_{k,t}^{-i}\}_t^{T} + \alpha)}$
Gibbs sampler can be generically derived (Heinrich 2009)

Typical case: Quotients of factors over levels $\ell$:

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

$n_{k,t}$ = count of co-occurrences between input and output values of a level (components and samples)

More complex variants covered by $q(k, t) \triangleq \frac{\text{beta}(\{n_{k,t}\}_{i=1}^T + \alpha)}{\text{beta}(\{n_{k,t}^{-i}\}_{i=1}^T + \alpha)}$
Gibbs sampler can be generically derived (Heinrich 2009)

Typical case: Quotients of factors over levels $\ell$:

$$
p(H_i|H_{\neg i}, V, A) \propto \prod_{\ell} \left[ \frac{n_{k,t}^{\neg i} + \alpha}{\sum_t n_{k,t}^{\neg i} + \alpha} \right]^{[\ell]} = \prod_{\ell} \left[ q(k, t) \right]^{[\ell]}
$$

$n_{k,t} = \text{count of co-occurrences between input and output values of a level (components and samples)}$

More complex variants covered by $q(k, t) \overset{\Delta}{=} \frac{\text{beta}({\{n_{k,t}\}}_{t=1}^{T} + \alpha)}{\text{beta}({\{n_{k,t}^{\neg i}\}}_{t=1}^{T} + \alpha)}$
Typology of NoMM substructures

- **N1. Dirichlet–multinomial**
  \[
  q(a, z) \, q(z, b)
  \]

- **E2. Autonomous edges**
  \[
  q(a, x \oplus y) \, q(x, b) \, q(y, c)
  \]

- **C2. Combined indices**
  \[
  q(a, x) \, q(b, y) \, q(k, c), \; k = f(x, y)
  \]

- **N2. Observed parameters**
  \[
  \theta_{a,z}^c \, q(z, b)
  \]

- **E3. Coupled edges**
  \[
  q(a, z) \, q(z, b) \, q(z, c)
  \]

- **C3. Interleaved indices**
  \[
  \approx q(a, z^1) \, q(b, z^2) \, q(z^1, c \oplus \tilde{c}) \, q(z^2, \tilde{c} \oplus c)
  \]

- **NoMM substructures:** Nodes, edges/branches, component indices/merging of edges:
  - Representation via \(q\)-functions and likelihood
  - Multiple samples per data point: \(q(a, x \oplus y)\) for respective level
  - “Library” incl. additional structures: alternative distributions, regression, aggregation etc. \(\leftrightarrow q\)-functions + other factors
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
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;
}.
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)
                    / (nkws[ksel] + betasum);
        psum += pp[hx][hy];
    } // for h
} // for h

Coding hints:
- Use variables to keep track of edge counts and states.
- Avoid inline comments for clarity.
- Use consistent indentation and spacing.

// 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)
                    / (nkws[ksel] + betasum);
        psum += pp[hx][hy];
    } // for h
} // for h

Gregor Heinrich

A generic approach to topic models 22/35
## Iteration 1

Document–topic matrix $\Theta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 5

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 10

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Iteration 15

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Iteration 20

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 30

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Iteration 40

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 50

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Iteration 60

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 80

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 100

Document–topic matrix $\phi$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 120

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 150

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 200

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 300

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Document–Topic distribution in Gibbs sampler

Iteration 500, converged ↔ stationary state

Document–topic matrix $\vartheta$ (200 documents, 50 topics)
Serial and parallel scaling methods:
- Generalised results for LDA to generic NoMMs, specifically (Porteous et al. 2008; Newman et al. 2009) + novel approach
- Problem: Sampling space for stat. dependent variables: $K \times L \times \ldots$
  → Independence assumption: Separate samplers with dimensions $K + L + \ldots \ll K \times L \times \ldots$
  - Empirical result: Iterations ↑, but topic quality comparable
  → Hybrid approaches with independent samplers highly effective
- Implementation: complexity covered by meta-sampler

### Table

<table>
<thead>
<tr>
<th>Model</th>
<th>Dim.</th>
<th>Serial Par.</th>
<th>Independence</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>500</td>
<td>✓</td>
<td>✓</td>
<td>30.2</td>
</tr>
<tr>
<td>PAM4</td>
<td>$40 \times 40$</td>
<td>✓</td>
<td>✓</td>
<td>7.4</td>
</tr>
<tr>
<td>PAM4</td>
<td>$40 \times 40$</td>
<td>✓</td>
<td>✓</td>
<td>24.1</td>
</tr>
<tr>
<td>PAM4</td>
<td>$40 \times 40$</td>
<td>✓</td>
<td>✓</td>
<td>49.8</td>
</tr>
</tbody>
</table>
Overview

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook
“How can generic models be applied to data in virtual communities?”
Typology → “Library” of NoMM substructures
→ Idea: Construct models from simple substructures that connect terminal nodes:
  ● Terminal nodes ↔ multimodal data (virtual communities...)
  ● Substructures ↔ relationships in data; latent semantics
→ Process:
  ● Assumptions on dependencies in data
  ● Iterative association to structures in model (usage of typology)
    ● Gibbs distribution known! ↔ model behaviour: \( q(x, y) = \text{“rich get richer”} \)
  ● Implementation and test with Gibbs meta-sampler; possibly iteration
Typology → “Library” of NoMM substructures
→ Idea: Construct models from simple substructures that connect terminal nodes:
  • Terminal nodes ↔ multimodal data (virtual communities...)
  • Substructures ↔ relationships in data; latent semantics
→ Process:
  • Assumptions on dependencies in data
  • Iterative association to structures in model (usage of typology)
    • Gibbs distribution known! ↔ model behaviour: $q(x, y) = \text{“rich get richer”}$
  • Implementation and test with Gibbs meta-sampler; possibly iteration
Application: Expert finding with tag annotations

- Scenario: Expert finding via documents with tag annotations
  - Authors of relevant documents → experts
  - Frequently documents with additional annotations, here: tags
  → Goal: Enable tag queries, improve quality of text queries
  - Problem: Tags often incomplete, partly wrong
  → Connection of tags and experts via topics

1. Data: For each document $m$: text $\tilde{\text{w}}_m$, authors $\tilde{\text{a}}_m$, tags $\tilde{\text{c}}_m$
2. Goal: Tag query $\tilde{\text{c}}'$: $p(\tilde{\text{c}}' | a) = \max$
   Word query $\tilde{\text{w}}': p(\tilde{\text{w}}' | a) = \max$
3. Terminal nodes: Authors in input, words and tags in output
Application: Expert finding with tag annotations

Scenario: Expert finding via documents with tag annotations
- Authors of relevant documents → experts
- Frequently documents with additional annotations, here: tags
  → Goal: Enable tag queries, improve quality of text queries
- Problem: Tags often incomplete, partly wrong
  → Connection of tags and experts via topics

(1) Data: For each document $m$: text $\vec{w}_m$, authors $\vec{a}_m$, tags $\vec{c}_m$
(2) Goal: Tag query $\vec{c}'$: $p(\vec{c}'|a) = \max$
   word query $\vec{w}'$: $p(\vec{w}'|a) = \max$
(3) Terminal nodes: Authors in input, words and tags in output
Scenario: Expert finding via documents with tag annotations

- Authors of relevant documents → experts
- Frequently documents with additional annotations, here: tags

→ Goal: Enable tag queries, improve quality of text queries

- Problem: Tags often incomplete, partly wrong
  → Connection of tags and experts via topics

(1) Data: For each document $m$: text $\vec{w}_m$, authors $\vec{a}_m$, tags $\vec{c}_m$

(2) Goal: Tag query $\vec{c}': p(\vec{c}' | a) = \max$
  word query $\vec{w}': p(\vec{w}' | a) = \max$

(3) Terminal nodes: Authors in input, words and tags in output
Scenario: Expert finding via documents with tag annotations

- Authors of relevant documents → experts
- Frequently documents with additional annotations, here: tags

→ Goal: Enable tag queries, improve quality of text queries
- Problem: Tags often incomplete, partly wrong
  → Connection of tags and experts via topics

(1) Data: For each document \( m \): text \( \vec{w}_m \), authors \( \vec{a}_m \), tags \( \vec{c}_m \)

(2) Goal: Tag query \( \vec{c}' \): \( p(\vec{c}'|a) = \max \)  
    word query \( \vec{w}' \): \( p(\vec{w}'|a) = \max \)

(3) Terminal nodes: Authors in input, words and tags in output
(4) Model assumptions:

(a) Expertise of an author is weighted with the portion of authorship
(b) Semantics of expertise expressed by topics $\mathbf{z}$. Each author has a single field of expertise (topic distribution).
(c) Semantics of tags expressed by topics $\mathbf{y}$
(4) Model assumptions:

(a) Expertise of an author is weighted with the portion of authorship
(b) Semantics of expertise expressed by topics $\mathbf{z}$. Each author has a single field of expertise (topic distribution).
(c) Semantics of tags expressed by topics $\mathbf{y}$.
(4) Model assumptions:
(a) Expertise of an author is weighted with the portion of authorship
(b) Semantics of expertise expressed by topics $z$. Each author has a single field of expertise (topic distribution).
(c) Semantics of tags expressed by topics $y$
(5) Model construction: (a) Start with terminal nodes (from step 3)
Model construction

\[ p(x, \ldots | \cdot) \propto a_{m,x} \cdot q(x, \ldots) \ldots \]

(b) Authorship \( \tilde{a}_m \) given as observed distribution
\( \rightarrow \) node samples author \( x \) of a word
Model construction

\[ \vec{c}_m \]

(c) Each author has only a single field of expertise (topic distribution) → \( q(x, z) \) associates (word-)topics with sampled authors \( x \) (cf. ATM)
Model construction

\[ p(x, z, \ldots \mid \cdot) \propto a_{m,x} q(x, z) \cdot q(z, w) \ldots \]

(d) Topic distribution over terms
\[ \rightarrow \text{connect } z \text{ and } w \text{ via } q(z, w) \]
Model construction

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

(e) Introduce tag topics \( y_{m,j} \) for \( c_{m,j} \) as distributions over tags
\( q(x, z \oplus y) \) overlays values for \( z \) and \( y \)
Model construction

\[ p(x, z, y \mid \cdot) \propto a_{m,x} \, q(x, z \oplus y) \, q(z, w) \, q(y, c) \]

(e) Introduce tag topics \( y_{m,j} \) for \( c_{m,j} \) as distributions over tags \( q(x, z \oplus y) \) overlays values for \( z \) and \( y \)
(e) Introduce tag topics $y_{m,j}$ for $c_{m,j}$ as distributions over tags $q(x, z \oplus y)$ overlays values for $z$ and $y$
Model construction – ordinary approach

**Expert–tag–topic model**

(Heinrich 2011)

\[
p(\bar{v}, \bar{d}, \bar{x}, \bar{z}, \bar{\phi}, \beta, \gamma) = p(\bar{v}|\bar{d}, \beta, \gamma) p(\bar{d}|\bar{x}, \gamma) \prod_{i=1}^{N} p(z_{m,n}|x_{m,n}) p(x_{m,n} | a_{m,n}, \alpha)
\]

\[
= \prod_{m=1}^{M} \prod_{n=1}^{N_m} \left( \prod_{j=1}^{J_m} p(w_{m,n}|z_{m,n}) p(z_{m,n} | x_{m,n}) p(x_{m,n} | a_{m,n}, \alpha) \right)
\]

\[
\cdot \prod_{k=1}^{K} p(c_{m,n}|z_{m,n}) p(y_{m,n} | c_{m,n}) p(c_{m,n} | a_{m,n}, \beta) \cdot p(\gamma) \cdot p(\Phi|\beta) \cdot p(\Psi|\beta)
\]

(3.3)
NIPS Corpus: 2.3 million words, 2037 authors, 165 tags

Retrieval: Average Precision @10:
- Term queries: ETT > ATM
- Tag queries: Similarly good AP values

Topic coherence (Mimno et al. 2011): ETT > ATM

Semi-supervised learning: Tag queries retrieve items without tags
Overview

- Introduction
- Generic topic models
- Inference methods
- Application to virtual communities
- Conclusions and outlook
Conclusions: Research contributions

- **Networks of Mixed Membership**: Generic model and domain-specific compact representation of topic models
- Inference algorithms: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
- Design process based on typology of NoMM substructures
- Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
- Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios
Conclusions: Research contributions

- **Networks of Mixed Membership**: Generic model and domain-specific compact representation of topic models
- Inference algorithms: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
  + Variational Inference for NoMMs (Heinrich and Goesele 2009)
- Design process based on typology of NoMM substructures
- Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents

∑ Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios
Conclusions: Research contributions

- **Networks of Mixed Membership**: Generic model and domain-specific compact representation of topic models
- Inference algorithms: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
  - Variational Inference for NoMMs (Heinrich and Goesele 2009)
- Design process based on typology of NoMM substructures
  - AMQ model: Meta-model for virtual communities as formal basis for scenario modelling (Heinrich 2010)
- Application to virtual communities: Expert–tag–topic model for expert finding with annotated documents
  \[ \sum \]
- Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios
Conclusions: Research contributions

- **Networks of Mixed Membership**: Generic model and domain-specific compact representation of topic models
- **Inference algorithms**: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
  - Variational Inference for NoMMs (Heinrich and Goesele 2009)
- **Design process based on typology of NoMM substructures**
  - AMQ model: Meta-model for virtual communities as formal basis for scenario modelling (Heinrich 2010)
- **Application to virtual communities**: Expert–tag–topic model for expert finding with annotated documents
  - Models ETT2 and ETT3 incl. novel NoMM structure; retrieval approaches (Heinrich 2011b)
- **Summarized contribution**: Contributes “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios
Conclusions: Research contributions

- **Networks of Mixed Membership**: Generic model and domain-specific compact representation of topic models
- **Inference algorithms**: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
  - Variational Inference for NoMMs (Heinrich and Goesele 2009)
- **Design process based on typology of NoMM substructures**
  - AMQ model: Meta-model for virtual communities as formal basis for scenario modelling (Heinrich 2010)
- **Application to virtual communities**: Expert–tag–topic model for expert finding with annotated documents
  - Models ETT2 and ETT3 incl. novel NoMM structure; retrieval approaches (Heinrich 2011b)
  - Graph-based expert search using ETT models: Integration of explorative search/browsing and distributions from topic models
- **Contribution to facilitated “model-based” construction of topic models**, specifically for virtual communities and other multimodal scenarios
Conclusions: Research contributions

Figure: ETT1: Expert search in community browser
Conclusions: Research contributions

- **Networks of Mixed Membership**: Generic model and domain-specific compact representation of topic models
- **Inference algorithms**: Generic Gibbs sampler
  - Fast sampling methods (serial, parallel, independent)
  - Implementation in Gibbs meta-sampler
    + Variational Inference for NoMMs (Heinrich and Goesele 2009)
- **Design process based on typology of NoMM substructures**
  + AMQ model: Meta-model for virtual communities as formal basis for scenario modelling (Heinrich 2010)
- **Application to virtual communities**: Expert–tag–topic model for expert finding with annotated documents
  + Models ETT2 and ETT3 incl. novel NoMM structure; retrieval approaches (Heinrich 2011b)
  + Graph-based expert search using ETT models: Integration of explorative search/browsing and distributions from topic models
- **Contribution to facilitated “model-based” construction of topic models, specifically for virtual communities and other multimodal scenarios**
Outlook

- New applications and NoMM structures, e.g., time as variable
- Alternative inference methods:
  - Generic Collapsed Variational Bayes (Teh et al. 2007): Structure similar to Collapsed Gibbs-Sampler
  - Non-parametric methods: Learning model dimensions using Dirichlet or Pitman–Yor process priors (Teh et al. 2004; Buntine and Hutter 2010), NoMM polymorphism (Heinrich 2011a)
- Improved support in design process:
  - Data-driven design: Search over model structures to obtain best model for data set
  - Architecture-specific Gibbs meta-sampler, e.g., massively-parallel or FPGA, cf. (Heinrich et al. 2011)
- Integration with interactive user interfaces: Models can be created on the fly, e.g., for visual analytics
Thank you!

Q + A
References


### Example: Text mining for semantic clusters

| Topic label       | Dominant terms according to $\varphi_{k,t} = p(\text{term} | \text{topic})$ |
|-------------------|--------------------------------------------------------------------------------|
| 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 Russlands 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 from 18400 dpa news messages, Jan. 2000 (Heinrich et al. 2005)
parameters $\vartheta +$ hyperparameters $\alpha \Leftrightarrow$ nodes $(\vartheta | \alpha)$

variables $k_i, x_i \Leftrightarrow$ edges $k_i, x_i$

plates (i.i.d. repetitions) $i, k \Leftrightarrow$ indexes $i +$ dimensions $k$
NoMM representation: Variable dependencies

\[ \alpha \vec{\vartheta}_k \vec{\vartheta}_k \vec{\vartheta}_k \]

\[ \ell = 1 \quad \ell = 2 \quad \ell = 3 \quad \ell = 4 \quad \ell = 5 \quad \ell = 6 \quad \ell = 7 \quad \ell = 8 \]

\[ \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \]

\[ \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \]

\[ X \quad \Theta \quad A \]

\[ h_i \quad h_i \quad v_i \quad h_i \quad v_i \]

\[ k_i^1 \quad k_i^2 \quad h_i^1 \quad h_i^2 \quad h_i^3 \quad h_i^4 \quad h_i^5 \quad h_i^6 \quad h_i^7 \quad h_i^8 \]

\[ \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \]

\[ v_i \quad v_i \quad v_i \quad v_i \quad v_i \quad v_i \quad v_i \quad v_i \quad v_i \quad v_i \]

\[ h_i \quad v_i \quad h_i \quad v_i \quad h_i \quad v_i \quad h_i \quad v_i \quad h_i \quad v_i \]

\[ \ell = 1 \quad \ell = 2 \quad \ell = 3 \quad \ell = 4 \quad \ell = 5 \quad \ell = 6 \quad \ell = 7 \quad \ell = 8 \]

\[ \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \quad \vec{\vartheta}_k \]

\[ \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \quad \alpha \]
**Collapsed Gibbs sampler**: Stochastic EM / MCMC:

- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

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

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler: Stochastic EM / MCMC:
- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:
  \[ H_i \sim p(H_i | H_{\neg i}, V, A). \] (1)
- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
**Collapsed Gibbs sampler**

- **Collapsed Gibbs sampler:** Stochastic EM / MCMC:
  - NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
  - For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:
    \[
    H_i \sim p(H_i \mid H_{\neg i}, V, A). 
    \]  
  - Stationary state: *full conditional* distribution (1) simulates posterior
  - Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler:

Collapsing Gibbs sampler: Stochastic EM / MCMC:
- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

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

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler: Stochastic EM / MCMC:

- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{-i}$, and observed, $V$:

$$H_i \sim p(H_i | H_{-i}, V, A).$$  \hspace{1cm} (1)

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsible Gibbs sampler

**Collapsed** Gibbs sampler: Stochastic EM / MCMC:
- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

\[
H_i \sim p(H_i | H_{\neg i}, V, A).
\]  

- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler

- Collapsed Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
  - For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

\[
H_i \sim p(H_i \mid H_{\neg i}, V, A).
\]  

  - Stationary state: full conditional distribution (1) simulates posterior
  - Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
**Collapsed Gibbs sampler**

- **Collapsed** Gibbs sampler: Stochastic EM / MCMC:
  - NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
  - For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:
    \[
    H_i \sim p(H_i | H_{\neg i}, V, A). \tag{1}
    \]
  - Stationary state: *full conditional* distribution (1) simulates posterior
  - Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
**Collapsed Gibbs sampler**

- **Collapsed Gibbs sampler**: Stochastic EM / MCMC:
  - NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
  - For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

    \[
    H_i \sim p(H_i | H_{\neg i}, V, A). \tag{1}
    \]

  - Stationary state: *full conditional* distribution (1) simulates posterior
  - Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
\textbf{Collapsed Gibbs sampler:} Stochastic EM / MCMC:

- NoMMs: parameters \( \Theta \) correlate with \( H \) \( \rightarrow \) marginalise
- For each data point, \( i \): draw latent variables, \( H_i = (y_i, z_i, \ldots) \), given all other data, latent, \( H_{\neg i} \), and observed, \( V \):

\[
H_i \sim p(H_i | H_{\neg i}, V, A).
\]

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler: Stochastic EM / MCMC:

- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

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

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
**Collapsed Gibbs sampler**: Stochastic EM / MCMC:

- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{-i}$, and observed, $V$:

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

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler: Stochastic EM / MCMC:
- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

$$H_i \sim p(H_i \mid H_{\neg i}, V, A).$$  \hspace{1cm} (1)

- Stationary state: *full conditional* distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Collapsed Gibbs sampler: Stochastic EM / MCMC:

- NoMMs: parameters $\Theta$ correlate with $H \rightarrow$ marginalise
- For each data point, $i$: draw latent variables, $H_i = (y_i, z_i, \ldots)$, given all other data, latent, $H_{\neg i}$, and observed, $V$:

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

- Stationary state: full conditional distribution (1) simulates posterior
- Faster absolute convergence for NoMMs than, e.g., variational inference (Heinrich and Goesele 2009)
Generative process on level $\ell$:

\[
x_i \sim \text{Mult}(x_i | \tilde{\vartheta}_k) \quad \tilde{\vartheta}_k \sim \text{Dir}(\tilde{\vartheta}_k | \tilde{\alpha}_j)
\]

\[
k = f_k(\text{parents}(x_i), i) \quad j = f_j(\text{known_parents}(x_i), i)
\]
Likelihood of all hidden and visible data \( X = \{H, V\} \) and parameters \( \Theta \):

\[
p(X, \Theta|A) = \prod_{\ell \in L} \left[ \prod_{i} p(x_{i,\text{out}}|\vec{\theta}_{x_{i,\text{in}}}) \cdot \prod_{k} p(\vec{\theta}_{k}|\vec{\alpha}) \right]^{[\ell]}
\]

- Data items \( \sim \text{Discrete} \)
- Components \( \sim \text{Dirichlet} \)
- Levels

\[
= \prod_{\ell \in L} \left[ \prod_{k} f(\vec{\theta}_k, \vec{n}_k, \vec{\alpha}) \right]^{[\ell]} \quad \vec{n}_k = (n_{k,1}, n_{k,2}, \cdots) \quad (4)
\]

- Product dependent on co-occurrences \( n_{k,t} \) between input and output values, \( x_{i,\text{in}}=k \) and \( x_{i,\text{out}}=t \), on each level \( \ell \)
- There are variants to component selection \( x_{i,\text{in}}=k \)
- There are mixture node variants, e.g., observed components
The conjugacy between the multinomial and Dirichlet distributions of model levels leads to a simple complete-data likelihood:

\[
p(X, \Theta | A) = \prod_{\ell} \prod_{i} \text{Mult}(x_{i}^{\ell} | \Theta^{\ell}, k_{i}^{\ell}) \prod_{k} \text{Dir}(\vec{\vartheta}_{k}^{\ell} | \vec{\alpha}_{j}^{\ell}) \tag{5}
\]

\[
= \prod_{\ell} \left[ \prod_{i} \vartheta_{k_{i}x_{i}}^{\ell} \prod_{k} \frac{1}{B(\vec{\alpha}_{j})} \prod_{t} \vartheta_{k,x_{i}}^{\alpha_{j}-1} \right]^{\ell} \tag{6}
\]

\[
= \prod_{\ell} \left[ \prod_{k} \frac{1}{B(\vec{\alpha}_{j})} \prod_{t} \vartheta_{k,x_{i}}^{\alpha_{j}+n_{k,t}-1} \right]^{\ell} \tag{7}
\]

\[
= \prod_{\ell} \left[ \prod_{k} \frac{B(\vec{n}_{k} + \vec{\alpha}_{j})}{B(\vec{\alpha}_{j})} \text{Dir}(\vec{\vartheta}_{k} | \vec{n}_{k} + \vec{\alpha}_{j}) \right]^{\ell} \tag{8}
\]

where brackets $[ \cdot ]^{\ell}$ enclose a particular level $\ell$.

$n_{k,t}$ is how often $k$ and $t$ co-occur.
Gibbs full conditionals are derived for groups of dependent hidden edges, $H^d_i \in H^d \subset X$ and “surrounding” edges $S^d_i \in S^d$ considered observed. All tokens co-located with a particular observation: $X^d_i = \{H^d_i, S^d_i\}$. Full conditional via chain rule applied to (8) with $\Theta$ integrated out:

$$p(H^d_i \mid X \setminus H^d_i, A) = \frac{p(H^d_i, S^d_i \mid X \setminus \{H^d_i, S^d_i\}, A)}{p(S^d_i \mid X \setminus \{H^d_i, S^d_i\}, A)}$$  \hspace{1cm} (9)$$

$$\propto p(X^d_i \mid X \setminus X^d_i, A) = \frac{p(X \mid A)}{p(X \setminus X^d_i \mid A)}$$  \hspace{1cm} (10)$$

$$= \prod_{\ell} \left[ \prod_k \frac{B(\vec{n}_k + \vec{\alpha}_j)}{B(\vec{n}_k \setminus X^d_i + \vec{\alpha}_j)} \right]^{\ell}$$  \hspace{1cm} (11)$$

$$\propto \prod_{\ell \in \{H^d, S^d\}} \left[ \frac{B(\vec{n}_k + \vec{\alpha}_j)}{B(\vec{n}_k \setminus X^d_i + \vec{\alpha}_j)} \right]^{\ell}$$  \hspace{1cm} (12)$$
Inference: $q$-functions

\[
q(k, t) = \frac{B(\n_k + \alpha_j)}{B(\n_k \setminus x_i^d + \alpha_j)}
\]

\[
|x_i^d| = 1 \quad n_{k,t}^{-i} + \alpha
\sum_t n_{k,t}^{-i} + \alpha
\]

\[
|x_i^d| = 2 \quad n_{k,t} \setminus x_{i,1}^d + \alpha \quad n_{k,t} \setminus x_{i,2}^d + \alpha + \delta(x_{i,1}^d - x_{i,2}^d)
\sum_t n_{k,t} \setminus x_{i,1}^d + \alpha \quad \sum_t n_{k,t} \setminus x_{i,2}^d + \alpha + 1
\]

\[
\ldots
\]

\[
\frac{n_{k,t}^{-i} + \alpha}{\sum_t n_{k,t}^{-i} + \alpha}
\]
$q$-functions: Pólya urn and sampling weights

Figure: Pólya urn: sampling with over-replacement.

\[
q(k, t) \triangleq \frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_k^{-i} + \alpha)} \quad \text{if} \quad |t| = 1
\]

\[
q(k, t) = \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}
\]

\[
t = \{u,v\} \quad \Rightarrow \quad \frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v)
\]

\[
\ldots
\]
\( q(k, t) \triangleq \frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_k - i + \alpha)} \quad |t| = 1 = \frac{n_{k,t}^{-t_i} + \alpha}{n_{k}^{-t_i} + T\alpha} = \) smoothed ratio of occurrences

\[
\begin{align*}
q(k, \{u,v\}) &\triangleq \frac{n_{k,u}^{-u_i} + \alpha}{n_{k}^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_{k}^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v) \\
\ldots
\end{align*}
\]
\[ q(k, t) \triangleq \frac{B(\tilde{n}_k + \alpha)}{B(\tilde{n}_k^{-i} + \alpha)} \quad |t|=1 \quad \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences} \]

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

...
$q$-functions: Pólya urn and sampling weights

Figure: Pólya urn: sampling with over-replacement.

$$q(k, t) \triangleq \frac{B(\tilde{n}_k + \alpha)}{B(\tilde{n}_k^{-i} + \alpha)} = \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$

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

...
functions: Pólya urn and sampling weights

\[ q(k, t) \triangleq \frac{\text{B}(\hat{n}_k + \alpha)}{\text{B}(\hat{n}_k^{-i} + \alpha)} \quad |t| \leq 1 \]

\[ \frac{n_k^{-t_i} + \alpha}{n_k + T\alpha} = \text{smoothed ratio of occurrences} \]

\[ t = \{u, v\} \quad \frac{n_k^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_k^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v) \]

Figure: Pólya urn and discrete parameters.
$q$-functions: Pólya urn and sampling weights

$E\{\vec{\vartheta}_k\}$

Figure: Pólya urn and discrete parameters.

$$q(k, t) \triangleq \frac{B(\tilde{n}_k + \alpha)}{B(\tilde{n}_k^{-t_i} + \alpha)} \cdot \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$

$$\frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v)$$
functions: Pólya urn and sampling weights

\[
q(k, t) \triangleq \frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_i^{-} + \alpha)} \quad |t| = 1
\]

\[
= \frac{n_{k,t}^{-t_i} + \alpha}{n_{k}^{-t_i} + T\alpha}
\]

smoothed ratio of occurrences

\[
\frac{\vec{n}_{v_i} + \alpha + \delta(u - v)}{n_{k}^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v)
\]
$q$-functions: Pólya urn and sampling weights

Figure: Pólya urn and discrete parameters.

$$q(k, t) \triangleq \frac{\text{B}(\tilde{n}_k + \alpha)}{\text{B}(\tilde{n}_k^{-t_i} + \alpha)} = \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences}$$

$$\frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v)$$
\( q(k, t) \triangleq \frac{B(\vec{n}_k + \alpha)}{B(\vec{n}_k - i + \alpha)} \left| t \right| \equiv \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences} \)

\[
\begin{align*}
q(k, t) & \equiv \frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \cdot \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \\
& \triangleq q(k, u \oplus v)
\end{align*}
\]
$q$-functions: Pólya urn and sampling weights

Figure: Pólya urn and discrete parameters.

\[ q(k, t) \triangleq \frac{B(\hat{n}_k + \alpha)}{B(\hat{n}_k^{-i} + \alpha)} \]  
\[ \trianglerighteq \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T \alpha} = \text{smoothed ratio of occurrences} \]

\[ t=\{u,v\} \]
\[ \frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T \alpha} \]
\[ \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T \alpha + 1} \]

\[ \trianglerighteq q(k, u \oplus v) \]

\[ q(k, \oplus) \]
$q$-functions: Pólya urn and sampling weights

Figure: Pólya urn and discrete parameters.

\[ q(k, t) \triangleq \frac{B(\tilde{n}_k + \alpha)}{B(\tilde{n}_k^{-i} + \alpha)} \quad \text{for} \quad |t| = 1 \]

\[ \equiv \frac{n_{k,t}^{-t_i} + \alpha}{n_k^{-t_i} + T\alpha} = \text{smoothed ratio of occurrences} \]

\[ t = \{u, v\} \]

\[ \frac{n_{k,u}^{-u_i} + \alpha}{n_k^{-u_i} + T\alpha} \]

\[ \frac{n_{k,v}^{-v_i} + \alpha + \delta(u - v)}{n_k^{-v_i} + T\alpha + 1} \triangleq q(k, u \oplus v) \]

\[ q(k, \bigoplus \bigotimes \bigodot) \]
$q$-functions: Pólya urn and sampling weights

\[ q(k, t) \triangleq \frac{B(\tilde{n}_k + \alpha)}{B(\tilde{n}_k^- + \alpha)} \mid_{|t|=1} \frac{n_{k,t}^- + \alpha}{n_k^- + T\alpha} = \text{smoothed ratio of occurrences} \]

\[ \begin{align*}
&\{u,v\} = \frac{n_{k,u}^- + \alpha}{n_k^- + T\alpha} \\
&\frac{n_{k,v}^- + \alpha + \delta(u - v)}{n_k^- + T\alpha + 1} \\triangleq q(k, u \oplus v) \\
&\ldots
\end{align*} \]

$\Delta$-functions: Pólya urn and discrete parameters.

Figure: Pólya urn and discrete parameters.
NoMM substructure library: Gibbs weights and likelihood

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Structure diagram</th>
<th>Gibbs sampler weight ( w )</th>
<th>Likelihood ( p ) for single token ( i )</th>
<th>Modelled aspect, example models</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1, E1, C1</td>
<td>Dir-Mult nodes, unbranched</td>
<td>( a_t )</td>
<td>( w(a_t) = q(a_i, z_1</td>
<td>q(z_i, b) ) ( E1 ): ( q(a_i, z_1</td>
<td>q(z_i, b) ) )</td>
</tr>
<tr>
<td>N2</td>
<td>Observed parameters</td>
<td>( a_t )</td>
<td>( w(a_t) = q(a_i, z_1</td>
<td>q(z_i, b) ) ( E2 ): ( q(a_i, z_1</td>
<td>q(z_i, b) ) )</td>
</tr>
<tr>
<td>N3</td>
<td>Non-Dirichlet prior</td>
<td>( a_t )</td>
<td>( w(a_t) = q(a_i, z_1</td>
<td>q(z_i, b) ) ( E3 ): ( q(a_i, z_1</td>
<td>q(z_i, b) ) )</td>
</tr>
<tr>
<td>N5, E4</td>
<td>Aggregation</td>
<td>( a_t )</td>
<td>( w(a_t) = q(a_i, z_1</td>
<td>q(z_i, b) ) ( E4 ): ( q(a_i, z_1</td>
<td>q(z_i, b) ) )</td>
</tr>
<tr>
<td>E2</td>
<td>Autonomous edges</td>
<td>( a_t )</td>
<td>( w(a_t) = q(a_i, z_1</td>
<td>q(z_i, b) ) ( E2A ): ( q(a_i, z_1</td>
<td>q(z_i, b) ) ) ( E2B ): ( q(a_i, z_1</td>
</tr>
<tr>
<td>E3</td>
<td>Coupled edges</td>
<td>( a_t )</td>
<td>( w(a_t) = q(a_i, z_1</td>
<td>q(z_i, b) ) ( E3A ): ( q(a_i, z_1</td>
<td>q(z_i, b) ) ) ( E3B ): ( q(a_i, z_1</td>
</tr>
<tr>
<td>C2</td>
<td>Combined indices</td>
<td>( b_t )</td>
<td>( w(b_t) = q(b_i, y_i</td>
<td>q(y_i, c_i) ) ( C2A ): ( q(b_i, y_i</td>
<td>q(y_i, c_i) ) ) ( C2B ): ( q(b_i, y_i</td>
</tr>
<tr>
<td>C3</td>
<td>Interleaved indices</td>
<td>( b_t )</td>
<td>( w(b_t) = q(b_i, y_i</td>
<td>q(y_i, c_i) ) ( C3A ): ( q(b_i, y_i</td>
<td>q(y_i, c_i) ) ) ( C3B ): ( q(b_i, y_i</td>
</tr>
<tr>
<td>C4</td>
<td>Switch</td>
<td>( b_t )</td>
<td>( w(b_t) = q(b_i, z_1</td>
<td>q(z_i, c_i) ) ( C4A ): ( q(b_i, z_1</td>
<td>q(z_i, c_i) ) ) ( C4B ): ( q(b_i, z_1</td>
</tr>
<tr>
<td>C5</td>
<td>Node coupling</td>
<td>( b_t )</td>
<td>( w(b_t) = q(b_i, z_1</td>
<td>q(z_i, c_i) ) ( C5A ): ( q(b_i, z_1</td>
<td>q(z_i, c_i) ) ) ( C5B ): ( q(b_i, z_1</td>
</tr>
</tbody>
</table>
Gibbs meta-sampler: Java data structure

MixItem // interface: node or edge

// global id (= unique variable name)
~ name : String
// ≥2 edges: multiple inputs C2
// ≥2 nodes: merged inputs C3
~ parents : List<MixItem>
// ≥2 edges: indep. branches E2
// ≥2 nodes: coupled branches E3
~ children : List<MixItem>
// type of item: seq., topic, qfixed...
~ datatype : enum
// link type: C and E classifications
~ linktype : enum

MixNet // represents a NoMM

// nodes of the NoMM
~ nodes : List<MixNode>
// edges of the NoMM
~ edges : List<MixEdge>
// sequences of the NoMM
~ sequences : List<MixSequence>
// constants for the NoMM
~ constants : Map<String, String>

MixNode // NoMM node

// parameters \( \theta_{kt} \)
~ theta : Variable
// counts \( n_k, \sum n_k \)
~ ntheta, nthetasum : Variable
// hyperparameter \( \alpha \)
~ alpha : Variable

MixEdge // NoMM edge

// variable \( x_{nm} \)
~ x : Variable
// range of \( x, T \)
~ T : Expression
// E2 edge siblings, for \( \Delta(\cdot) \) expansion
~ siblingsE2 : List<MixEdge>
// flag: parent node emits subset of range
~ sparse : boolean

MixSequence // NoMM sequence

// subsequences, null for leaf
~ subseqs : List<MixSequence>
// supersequence, null for root
~ superseq : MixSequence
// sequence index variables, \( m, n, s \)
~ m, n, s : Variable
// sequence index ranges: \( M, N_m, W \)
~ M, Mq, Nm, Nmq, W, Wq : Expression
// flag: fixed topics for query
~ qfixed : boolean
/** run the main Gibbs sampling kernel */
public void run(int niter) {
  // iteration loop
  for (int iter = 0; iter < niter; iter++) {
    // major loop, sequence [m][n]
    for (int m = 0; m < M; m++) {
      // component selectors
      int mxsel = -1;
      int mxjsel = -1;
      int ksel = -1;

      // minor loop, sequence [m][n]
      for (int n = 0; n < w[m].length; n++) {
        double psum;
        double u;
        // decrement counts
        nmx[m][x[m][n]]--;
        nmxy[mxsel][y[m][n]]--;
        nmxysum[mxsel]--;
        if (x[m][n] == 0) ksel = 0;
        else if (y[m][n] == 0)
          ksel = 1 + x[m][n];
        else
          ksel = 1 + X + y[m][n];
        nkw[ksel][w[m][n]]--;
        nkwsum[ksel]--;

        // compute weights
        psum = 0;
        for (hx = 0; hx < X; hx++) {
          for (hy = 0; hy < Y; hy++) {
            // hidden edge
            psum += pp[hx][hy];
          }
        }

        // assign topics
        u = rand.nextDouble() * psum;
        psum = 0;
        SAMPLED:
        // each edge value
        for (hx = 0; hx < X; hx++) {
          for (hy = 0; hy < Y; hy++) {
            psum += pp[hx][hy];
            if (u <= psum)
              break SAMPLED;
          }
        }

        // increment counts
        nmx[m][x[m][n]]++;
        nmxy[mxsel][y[m][n]]++;
        nmxysum[mxsel]++;
        if (x[m][n] == 0) ksel = 0;
        else if (y[m][n] == 0)
          ksel = 1 + x[m][n];
        else
          ksel = 1 + X + y[m][n];
        nkw[ksel][w[m][n]]++;
        nkwsum[ksel]++;
      }
    }
  }
}

Figure: Generated Gibbs kernel for hPAM2 model in Fig. ??.

Gregor Heinrich
A generic approach to topic models 54 / 35
Idea: Exploit saliency of few elements → compute only largest (=most likely) weights

Approximate normalisation via vector norms (Porteous et al. 2008)

Generalisation to multiple dependent variables: more expensive higher-order vector norms ↔ higher sparsity of sampling space
Fast parallel sampling: Synchronisation methods

Multi-processor parallelisation using shared memory (OpenMP)

Main challenge: synchronisation and communication of global data

Synchronisation methods (LDA + generic NoMMs):

a. Naïve synchronisation locks
b. Query read-only $\varphi$ + MAP update step for $\varphi$ ("split-state")
c. Local copies $\varphi$ + reduction step (=AD-LDA (Newman et al. 2009))
Fast sampling: Serial × parallel

Figure: Speed-up for fast sampling methods: LDA.
Fast sampling: Serial $\times$ parallel $\times$ independent

Figure: Speed-up for combined fast samplers: PAM4 (2 dependent variables).
Fast sampling: The impact of assumed independence

**Figure:** Perplexity over iterations. Example model: PAM4.
ETT1 model: Derivation using NoMM structure

Lining up $q$-functions:

$$p(x, z, y | \cdot) \propto a_{m,x} q(x, z \oplus y) q(z, w) q(y, c)$$  \hspace{1cm} (13)

Transforming to standard Gibbs full conditionals:

$$p(x_{m,n} = x, z_{m,n} = z | \cdot) \propto a_{m,x} \cdot \frac{n^{-\{x,z\}_{m,n}} + \alpha}{n_x^{-\{x,z\}_{m,n}} + K\alpha} \cdot \frac{n^{-z_{m,n}} + \beta}{n_z^{-z_{m,n}} + V\beta}$$  \hspace{1cm} (14)

$$p(x_{m,j} = x, y_{m,j} = y | \cdot) \propto a_{m,x} \cdot \frac{n^{-\{x,y\}_{m,j}} + \alpha}{n_y^{-\{x,y\}_{m,j}} + K\alpha} \cdot \frac{n^{-y_{m,j}} + \gamma}{n_y^{-y_{m,j}} + C\gamma}$$  \hspace{1cm} (15)

Retrieval über Anfrage-Likelihood-Modell:

$$p(\vec{w} | a) = \prod_{w \in \vec{w}} \sum_{z} \vartheta_{a,z} \varphi_{z,w} \quad p(\vec{c} | a) = \prod_{c \in \vec{c}} \sum_{y} \vartheta_{a,y} \psi_{y,c}.$$  \hspace{1cm} (16)
ETT1 model: Derivation using NoMM structure

Lining up $q$-functions:

$$p(x, z, y \mid \cdot) \propto a_{m,x} \cdot q(x, z \oplus y) \cdot q(z, w) \cdot q(y, c)$$  \hspace{1cm} (13)

Transforming to standard Gibbs full conditionals:

$$p(x_{m,n}=x, z_{m,n}=z \mid \cdot) \propto a_{m,x} \cdot \frac{n^{-\{x,z\}_{m,n}} + \alpha}{n^{-\{x,z\}_{m,n}} + K\alpha} \cdot \frac{n^{-z_{m,n}} + \beta}{n^{-z_{m,n}} + V\beta}$$  \hspace{1cm} (14)

$$p(x_{m,j}=x, y_{m,j}=y \mid \cdot) \propto a_{m,x} \cdot \frac{n^{-\{x,y\}_{m,j}} + \alpha}{n^{-\{x,y\}_{m,j}} + K\alpha} \cdot \frac{n^{-y_{m,j}} + \gamma}{n^{-y_{m,j}} + C\gamma}$$  \hspace{1cm} (15)

Retrieval über Anfrage-Likelihood-Modell:

$$p(\bar{w} \mid a) = \prod_{w \in \bar{w}} \sum_{z} \vartheta_{a,z} \phi_{z,w}$$

$$p(\bar{c} \mid a) = \prod_{c \in \bar{c}} \sum_{y} \vartheta_{a,y} \psi_{y,c}.$$  \hspace{1cm} (16)
ETT1 model: Derivation using ordinary method (excerpt)

(Heinrich 2011b)

\[ p(\tilde{w}, \tilde{c}, \tilde{d}, \tilde{x}, \tilde{z}_x, \tilde{z}_\alpha, \tilde{z}_\beta, \tilde{z}_\gamma) = p(\tilde{w}|\tilde{c}, \tilde{d}, \tilde{x}, \Theta, \Phi) \cdot p(\tilde{c}|\tilde{d}, \tilde{x}, \Theta, \Psi) \cdot p(\tilde{x}|\tilde{z}_x, \Psi) \]

\[ = \prod_{m=1}^{M} \left( \prod_{n=1}^{N_m} p(w_{mn}|\tilde{c}_{mn}, \tilde{d}_{mn}) \cdot p(\tilde{c}_{mn}|\tilde{d}_{mn}, \alpha_{mn}) \cdot p(\tilde{d}_{mn}|\alpha_{mn}) \cdot p(\tilde{z}_x|\tilde{z}_\alpha, \Psi) \cdot p(\tilde{z}_\alpha|\Psi) \cdot p(\tilde{z}_\beta|\Psi) \cdot p(\tilde{z}_\gamma|\Psi) \right) \]  
\[ \cdot p(\tilde{x}|\tilde{z}_x, \Psi) \cdot p(\tilde{z}_x|\tilde{z}_\alpha, \Psi) \cdot p(\tilde{z}_\alpha|\Psi) \cdot p(\tilde{z}_\beta|\Psi) \cdot p(\tilde{z}_\gamma|\Psi) \hspace{1cm} (E.1) \] 

\[ \Rightarrow p(\tilde{w}, \tilde{c}, \tilde{d}, \tilde{x}, \tilde{z}_x, \tilde{z}_\alpha, \tilde{z}_\beta, \tilde{z}_\gamma) = \int \int \int \int \int \int \int \int p(\tilde{w}, \tilde{c}, \tilde{d}, \tilde{x}, \tilde{z}_x, \tilde{z}_\alpha, \tilde{z}_\beta, \tilde{z}_\gamma) \, d\tilde{w} \, d\tilde{c} \, d\tilde{d} \, d\tilde{x} \, d\tilde{z}_x \, d\tilde{z}_\alpha \, d\tilde{z}_\beta \, d\tilde{z}_\gamma \hspace{1cm} (E.2) \]

\[ p(z_i=k, x_i=x|w_i=t, \tilde{z}_{\tilde{x}i}, \tilde{z}_{\tilde{\alpha}i}, \tilde{z}_{\tilde{\beta}i}, \tilde{z}_{\tilde{\gamma}i}, \tilde{d}_i, \tilde{c}_i) \]

\[ = \frac{p(w_i, \tilde{c}_i, \tilde{d}_i, \tilde{x}_i, \tilde{z}_{\tilde{x}i}, \tilde{z}_{\tilde{\alpha}i}, \tilde{z}_{\tilde{\beta}i}, \tilde{z}_{\tilde{\gamma}i}, \tilde{d}_i, \tilde{c}_i)}{p(w_i, \tilde{c}_i, \tilde{d}_i, \tilde{x}_i, \tilde{z}_{\tilde{x}i}, \tilde{z}_{\tilde{\alpha}i}, \tilde{z}_{\tilde{\beta}i}, \tilde{z}_{\tilde{\gamma}i}, \tilde{d}_i, \tilde{c}_i) + \sum_{m \neq k} p(w_i, \tilde{c}_i, \tilde{d}_i, \tilde{x}_i, \tilde{z}_{\tilde{x}i}, \tilde{z}_{\tilde{\alpha}i}, \tilde{z}_{\tilde{\beta}i}, \tilde{z}_{\tilde{\gamma}i}, \tilde{d}_i, \tilde{c}_i)} \hspace{1cm} (E.3) \]
ETT1 evaluation: **Truncated Average Precision**

Truncated Average Precision (AP) is a measure used in information retrieval to evaluate the quality of a ranked list of retrieved documents. It is calculated as the average of the precision scores obtained at the top positions of the ranked list, truncated to a certain number of positions. The formula for AP@5 is shown below:

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

For AP@5, we assume there are 3 relevant documents in the corpus. The figure illustrates the calculation process for AP@5.

The figure shows: Average Precision at 5 (assuming 3 relevant documents in corpus)
ETT1 results: Term Retrieval

query: svm support vector machine | kernel classifier hyperplane regression

1. Scholkopf_B, lik = −76.272, tokens = 2830, docs = 10: judged relevant
   ✓ From Regularization Operators to Support Vector Kernels (9); Improving the Accuracy and Speed of Support Vector Machines (9); Shrinking the Tube: A New Support Vector Regression Algorithm (11) …

2. Smola_A, lik = −77.509, tokens = 2760, docs = 11: judged relevant
   ✓ Support Vector Regression Machines (9); Prior Knowledge in Support Vector Kernels (10); Support Vector Method for Novelty Detection (12) The Entropy Regularization Information Criterion (12, support vector machines, regularization) …

3. Vapnik_V, lik = −77.525, tokens = 2332, docs = 10: judged relevant
   ✓ Support Vector Regression Machines (9); Prior Knowledge in Support Vector Kernels (10); Prior Knowledge in Support Vector Kernels (10); Support Vector Method for Multivariate Density Estimation (12); …

4. Crisp_D, lik = −81.401, tokens = 699, docs = 2: judged relevant
   ✓ A Geometric Interpretation of t/-SVM Classifiers (12); Uniqueness of the SVM Solution (12)

5. Burges_C, lik = −81.630, tokens = 1309, docs = 5: judged relevant
   ✓ Improving the Accuracy and Speed of Support Vector Machines (9); A Geometric Interpretation of t/-SVM Classifiers (12); Uniqueness of the SVM Solution (12) …

6. Laskov_P, lik = −84.275, tokens = 738, docs = 1: judged relevant
   ✓ An Improved Decomposition Algorithm for Regression Support Vector Machines (12)

7. Steinhage_V, lik = −84.600, tokens = 438, docs = 1: judged irrelevant
   × Nonlinear Discriminant Analysis Using Kernel Functions (12)

8. Bennett_K, lik = −86.754, tokens = 384, docs = 1: judged relevant
   ✓ Semi-Supervised Support Vector Machines (11)

9. Herbrich_R, lik = −86.754, tokens = 462, docs = 2: judged irrelevant
   × Classification on Pairwise Proximity Data (11); Bayesian Transduction (12, classification)

10. Chapelle_O, lik = −87.431, tokens = 494, docs = 2: judged relevant
    ✓ Model Selection for Support Vector Machines (12); Transductive Inference for Estimating Values of Functions (12, regression, classification)
ETT1 results: Tag retrieval

query: face recognition

1. Movellan_J, lik = −4.680, tokens = 3153, docs = 8: judged relevant
✓ Dyn. Features for Visual Speechreading: A System Comparison (9, no tags); Image Representation for Facial Expression Coding (12, tags: face recognition, image, ICA); Visual Speech Recognition with Stochastic Networks (7, tags: HMM, speech recognition) . . .

2. Bartlett_M, lik = −4.951, tokens = 812, docs = 3: judged relevant
✓ Viewpoint Invariant Face Recognition using ICA and Attractor Networks (9, tags: face recognition, invariances, pattern recognition); Image Representation for Facial Expression Coding (12, tags: face recognition, image, ICA) . . .

3. Dailey_M, lik = −4.952, tokens = 903, docs = 2: judged relevant
✓ Task and Spatial Frequency Effects on Face Specialization (10, tags: face recognition); Facial Memory Is Kernel Density Estimation (Almost) (11, no tags)

4. Padgett_C, lik = −4.974, tokens = 499, docs = 1: judged relevant
✓ Representing Face Images for Emotion Classification (9, tags: classification, face recognition, image)

5. Hager_J, lik = −5.023, tokens = 377, docs = 2: judged relevant
✓ Classifying Facial Action (8, tags: classification); Image Representation for Facial Expression Coding (12, tags: face recognition, image, ICA)

6. Ekman_P, lik = −5.027, tokens = 374, docs = 2: judged relevant
✓ Image Representation for Facial Expression Coding (12, tags: face recognition, image, ICA); Classifying Facial Action (8, tags: classification)

7. Phillips_P, lik = −5.127, tokens = 795, docs = 1: judged relevant
✓ Support Vector Machines Applied to Face Recognition (11, tags: face recognition, SVM)

8. Gray_M, lik = −5.159, tokens = 470, docs = 2: judged irrelevant
× Dynamic Features for Visual Speechreading: A Systematic Comparison (9, text: dynamic visual features; no tags)

9. Lawrence_D, lik = −5.217, tokens = 265, docs = 1: judged relevant
✓ SEXNET: A Neural Network Identifies Sex From Human Faces (3, tags: neural networks, object recognition, pattern recognition)
ETT1 results: Tag query and expert topics

**Tag: face recognition (ETT1/J20)**

- 0.82702 face images faces image facial visual human video database detection
- 0.09392 image images texture pixel resolution pyramid regions pixels region search
- 0.02696 wavelet video view images tracking user camera image motion shape
- 0.00117 eeg brain ica artifacts subjects activity subject erp signals scalp
- 0.00100 image images visual vision optical pixel surface edge disparity receptive
- 0.00094 orientation cortical dominance ocular cortex development lateral eye cells visual
- 0.00084 hinton object image energy cost images code visible zemel codes

**Author: Movellan_J (ETT1/J20)**

- 0.53816: face images faces image facial visual human video database detection
- 0.16216: image images texture pixel resolution pyramid regions pixels region search
- 0.08954: speech speaker acoustic vowel phonetic phoneme utterances spoken formant
- 0.06216: bayesian prior density posterior entropy evidence likelihood distributions
- 0.03939: filter frequency signals phase channel amplitude frequencies temporal spectrum
- 0.03508: activation boltzmann annealing temperature neuron stochastic schedule machine
- 0.02770: cell firing cells neuron activity excitatory inhibitory synaptic potential membrane
- 0.02154: convergence stochastic descent optimization batch density global update

**Author: Cottrell_G (ETT1/J20)**

- 0.41865: recurrent nets correlation cascade activation connection epochs representations
- 0.27523: face images faces image facial visual human video database detection
- 0.17531: subjects human stimulus cue subject trials experiment perceptual psychophysical
- 0.11287: tangent transformation image simard images invariant invariance euclidean
- 0.07130: modules attractors cortex phase olfactory frequency bulb activity oscillatory eeg
- 0.06143: word connectionist representations words activation production cognitive musical
- 0.03695: node activation graph cycle nets message recurrence links connection child
- 0.02049: visual attention contour search selective orientation iiiii region saliency segment
ETT1 results: Topic coherence

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 2011b)

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

1. A. orientation  
   B. cortex  
   C. visual  
   D. ocular  
   E. acoustic  
   F. eye  

2. A. likelihood  
   B. mixture  
   C. theorem  
   D. density  
   E. em  
   F. prior  

3. A. risk  
   B. return  
   C. stock  
   D. trading  
   E. processor  
   F. prediction  

4. A. language  
   B. word  
   C. stress  
   D. grammar  
   E. neural  
   F. syllable  

5. A. circuit  
   B. bayesian  
   C. analog  
   D. voltage  
   E. vlsi  
   F. chip  

6. A. validation  
   B. set  
   C. variance  
   D. regression  
   E. selection  
   F. bias

(a) Topic intrusion experiment  
(b) Coherence scores
ETT1 results: Topic coherence

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 2011b)

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

(a) Topic intrusion experiment

(b) Coherence scores

(a) Topic intrusion experiment

(b) Coherence scores