Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Latente Dirichlet-Allokation

Tobias Scheffer
Peter Haider
Paul Prasse
Uwe Dick

Themenmodellierung

- Themenmodellierung (Topic modeling) liefert Methoden, große elektronische Archive automatisch zu organisieren, verstehen, durchsuchen und zusammenzufassen
  1. Versteckte Themenmuster finden, die in der Dokumentensammlung reflektiert sind
  2. Dokumente anhand der gefundenen Themen annotieren
  3. Annotationen verwenden, um Dokumente zu organisieren und durchsuchen

Themenmodellierung (Beispielsystem)

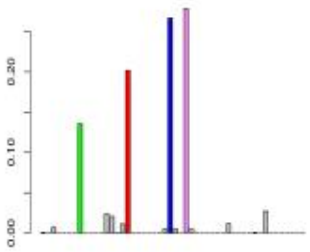
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal genera compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Probabilistische Modellierung

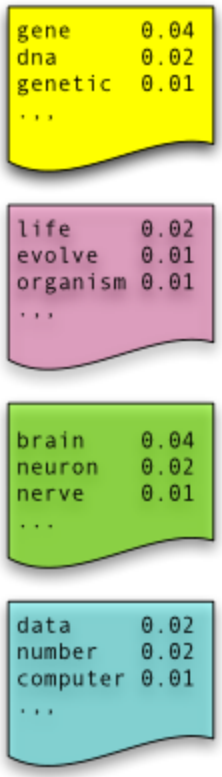
1. Behandlung der Daten als Beobachtungen, die aus einem generativen probabilistischen Prozess mit versteckten Variablen entstehen
   - Bei Dokumenten reflektieren die versteckten Variablen die thematische Struktur der Textsammlung
2. Versteckte Struktur finden mit Posterior-Inferenz
   - *Was sind die Themen, die diese Sammlung beschreiben?*
3. Neue Daten in das geschätzte Modell „einsortieren“
   - Wie passt das neue Dokument in die Themenstruktur?

Latente Dirichlet-Allokation

- Generatives Modell für Texte:
  - Modelliert gemeinsame Wahrscheinlichkeit für Beobachtungen, Label und versteckte Variablen.
  - Beschreibt den Prozess wie Beobachtungen, Label und versteckte Variablen erzeugt werden.
- Beobachtungen sind Wörter der Dokumente.
- Versteckte Variablen sind Themen für jedes Dokument und die Zusammensetzung der Themen an sich

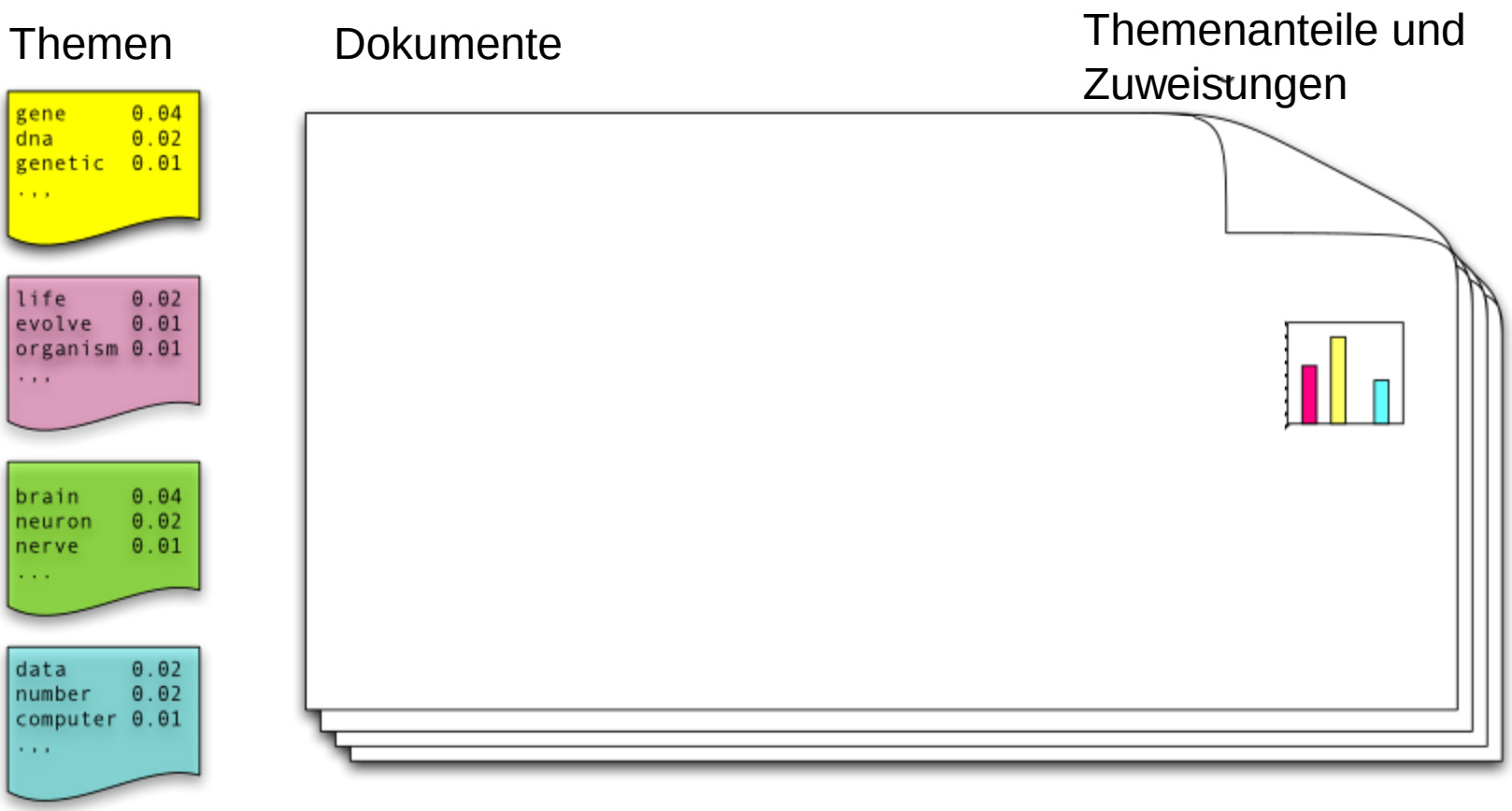
Generatives Modell

Themen



- Themen sind Verteilungen über die Wörter des Vokabulars

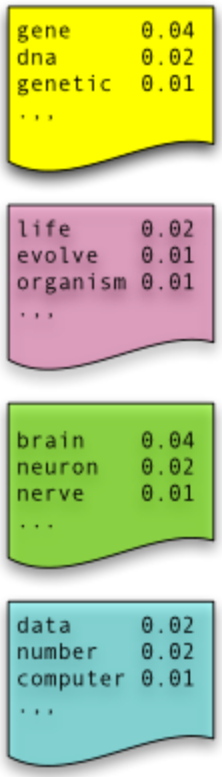
Generatives Modell



- Themen sind Verteilungen über die Wörter des Vokabulars
- Jedes Dokument ist eine Mischung aus (korpus-globalen) Themen

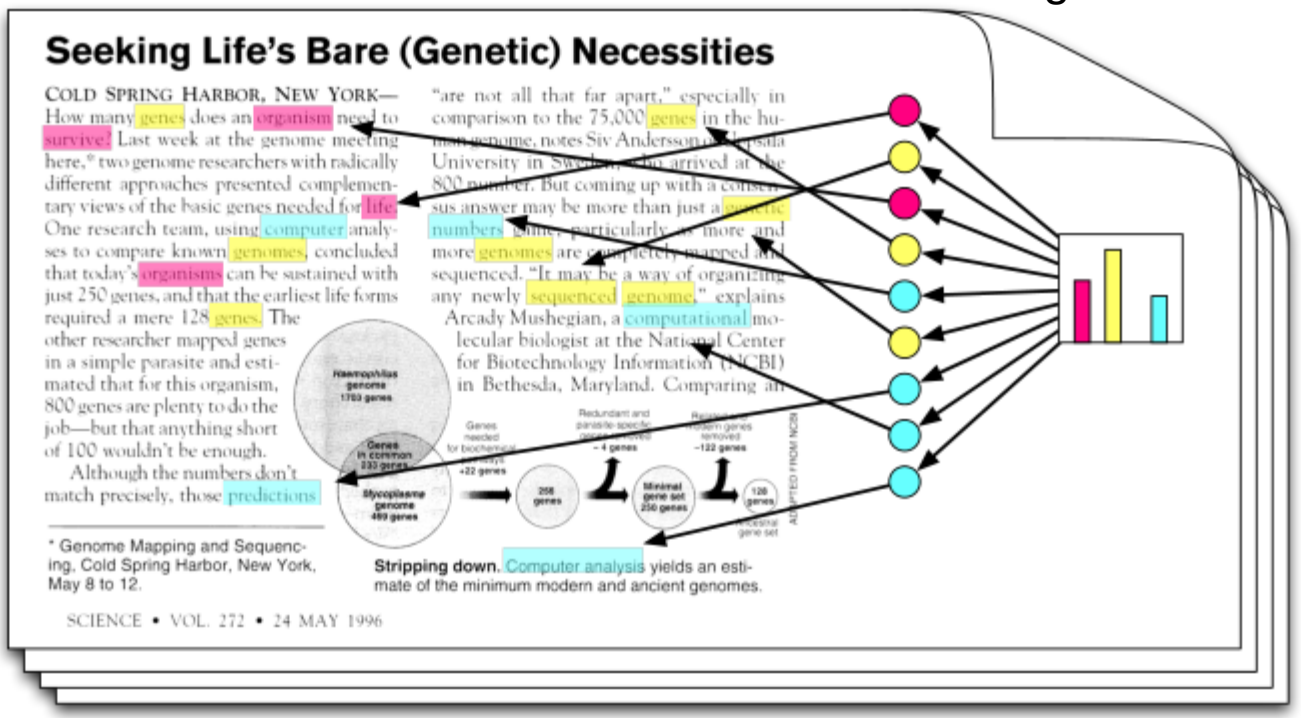
Generatives Modell

Themen



Dokumente

Themenanteile und Zuweisungen



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

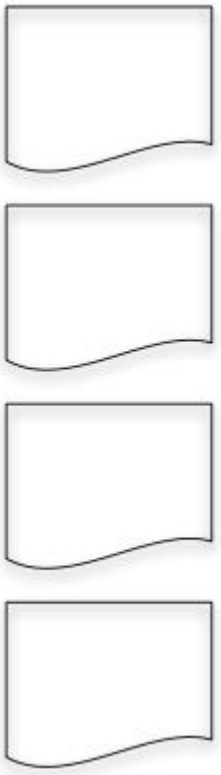
- Themen sind Verteilungen über die Wörter des Vokabulars
- Jedes Dokument ist eine Mischung aus (korpus-globalen) Themen
- Jedes Wort ist aus einem der Themen gezogen

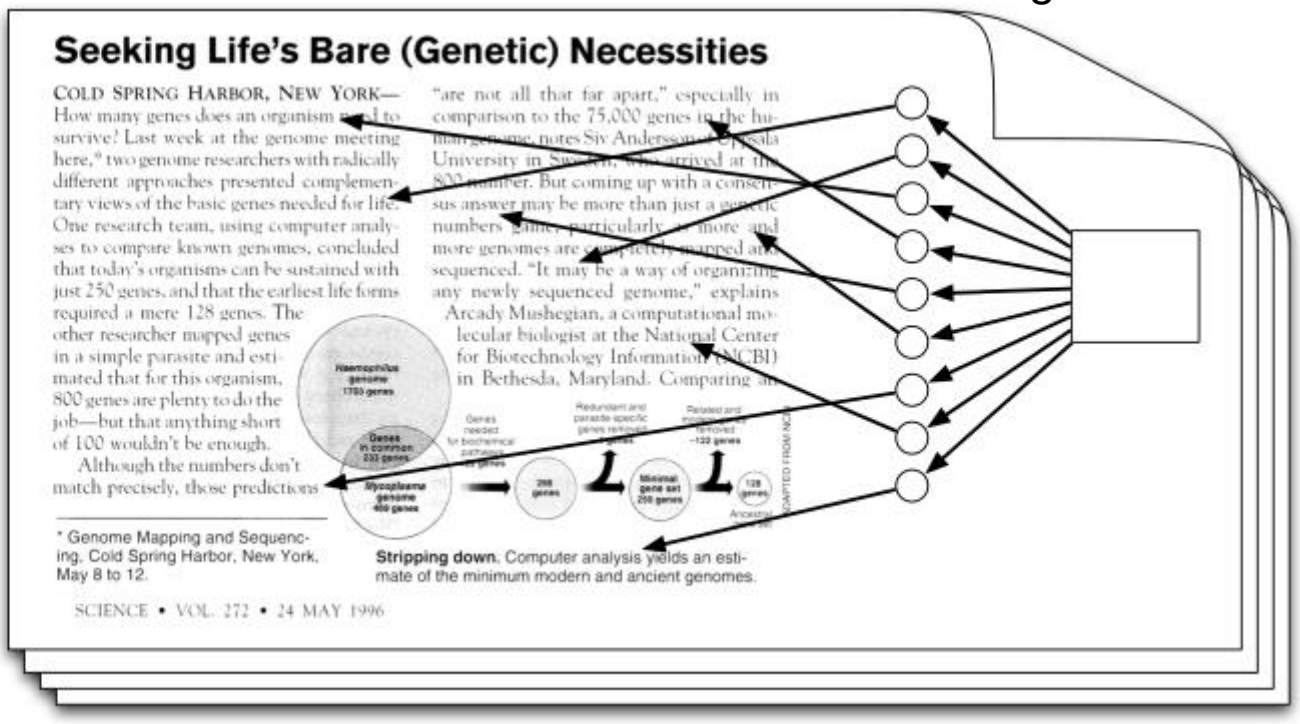
Inferenzproblem

Themen

Dokumente

Themenanteile und Zuweisungen





Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

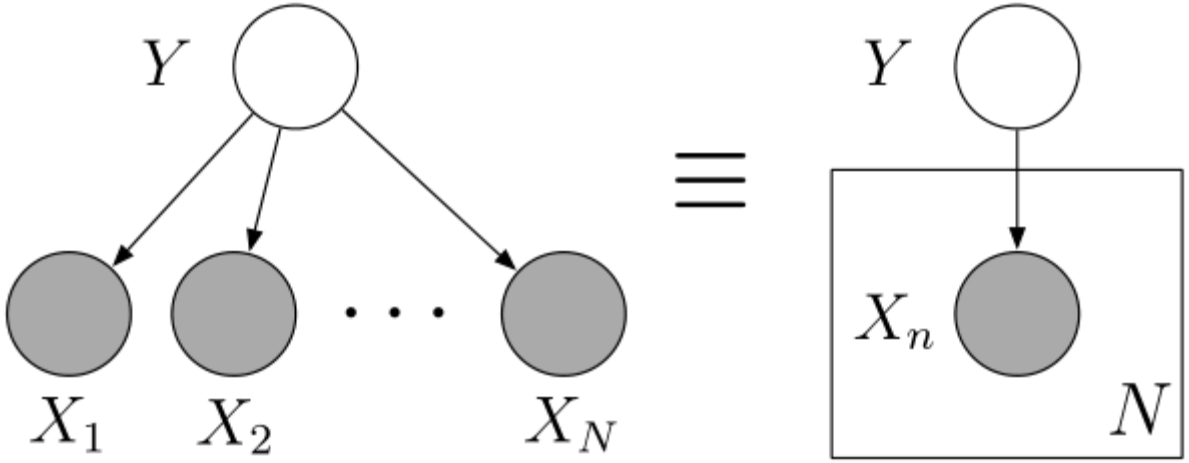
SCIENCE • VOL. 272 • 24 MAY 1996

- In Wirklichkeit sind nur die Wörter beobachtet
- Das Ziel ist, die zugrundeliegende Themenstruktur zu inferieren

Latente Dirichlet-Allokation

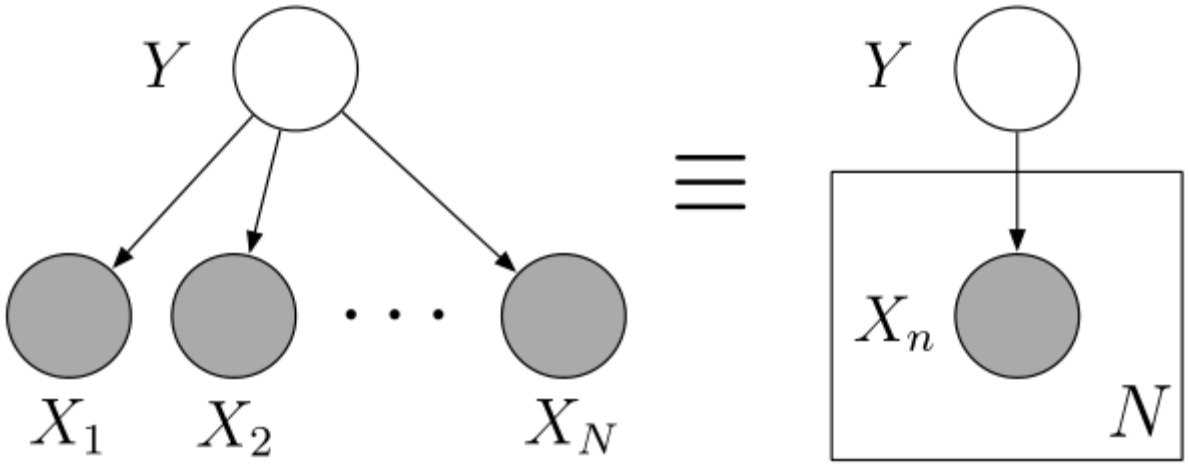
- Generatives Modell für Texte (als bag-of-words)
- Jeder Text ist Mischung verschiedener Themen:
  - Für jedes Thema ist die Zusammensetzung (die Worthäufigkeiten) eine versteckte Variable.
  - Für jeden Text ist die Mischung über die Themen eine versteckte Variable.
  - Wörter sind Beobachtungen des generativen Modells:
    - Jeder Stelle des Dokuments ist ein Thema gemäß der Themenmischung des Dokuments zugewiesen.
    - An jeder Stelle wird ein Wort erzeugt, gemäß den Worthäufigkeiten des entsprechenden Themas

Graphische Modelle



- Knoten sind Zufallsvariablen
- Kanten beschreiben mögliche Abhängigkeiten
- Beobachtete Variablen sind schattiert
- Tafeln (plates) beschreiben replizierte Struktur

Graphische Modelle



- Struktur des Graphen definiert bedingte Unabhängigkeiten zwischen den Zufallsvariablen
- Obiger Graph bedeutet:

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

Latente Dirichlet-Allokation

- Im Folgenden seien:
  - K die Anzahl der Themen,
  - V die Größe des Vokabulars,
  - D die Anzahl der Dokumente und
  - N_d die Anzahl von Wörtern im d-ten Dokument.

Latente Dirichlet-Allokation

- Für jedes Thema ist die Zusammensetzung (die Worthäufigkeiten) eine versteckte Variablen:
  - Für jedes Thema $k \in \{1,\ldots,K\}$ ist β_k eine V-dimensionale Zufallsvariable
  - Alle β_k besitzen die gleiche (Prior)-Verteilung $P(\beta_k \mid \eta)$
  - η ist Hyperparameter der gemeinsamen Verteilung (später mehr)

Latente Dirichlet-Allokation

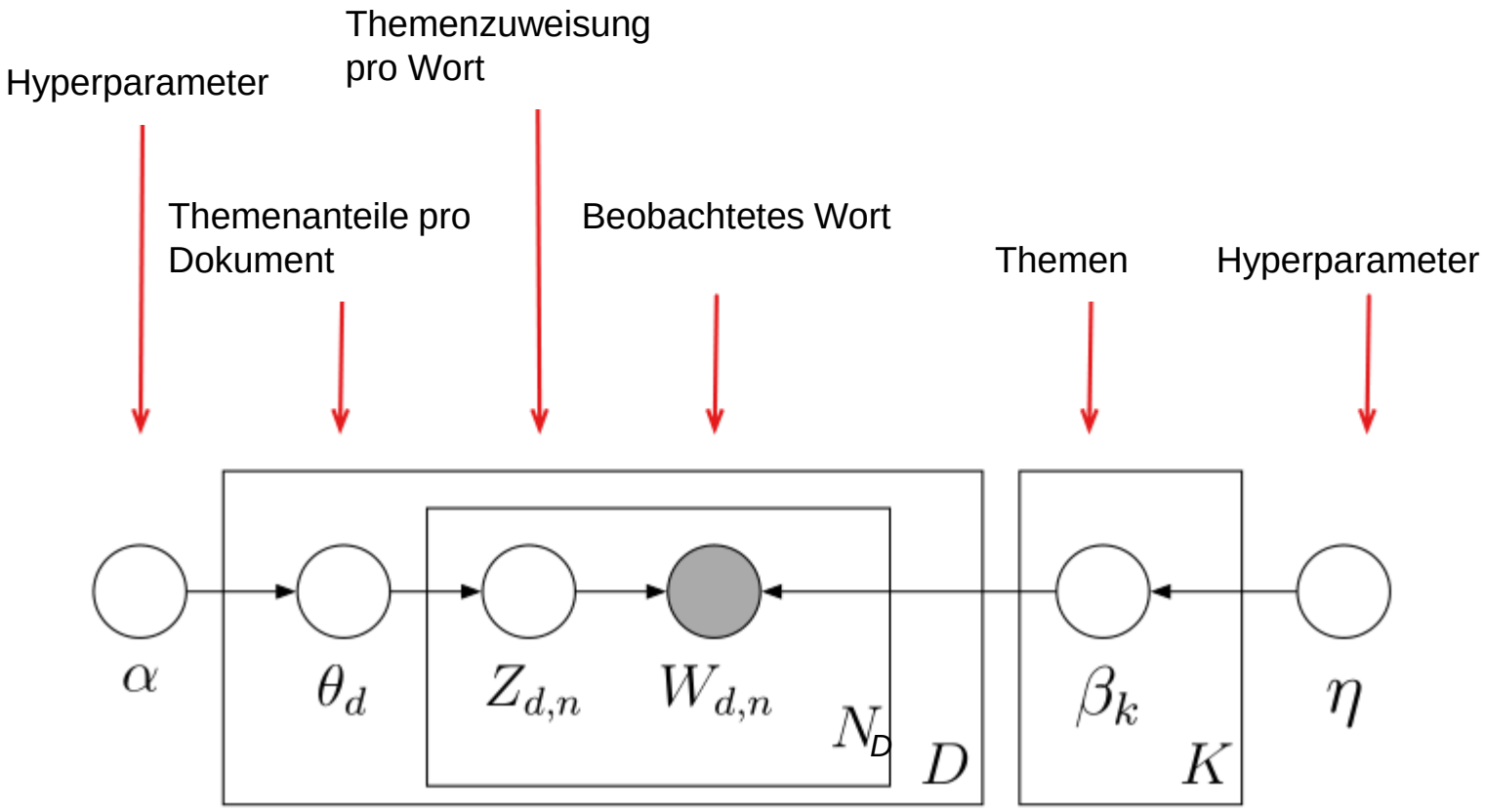
- Für jeden Text ist die Mischung über die Themen eine versteckte Variable:
  - Für jedes Dokument $d \in \{1,\ldots,D\}$ ist θ_d eine K-dimensionale Zufallsvariable
  - Alle θ_d besitzen die gleiche (Prior-)Verteilung $P(\theta_d \mid \alpha)$
  - α ist Hyperparameter der gemeinsamen Verteilung (später mehr)

Latente Dirichlet-Allokation

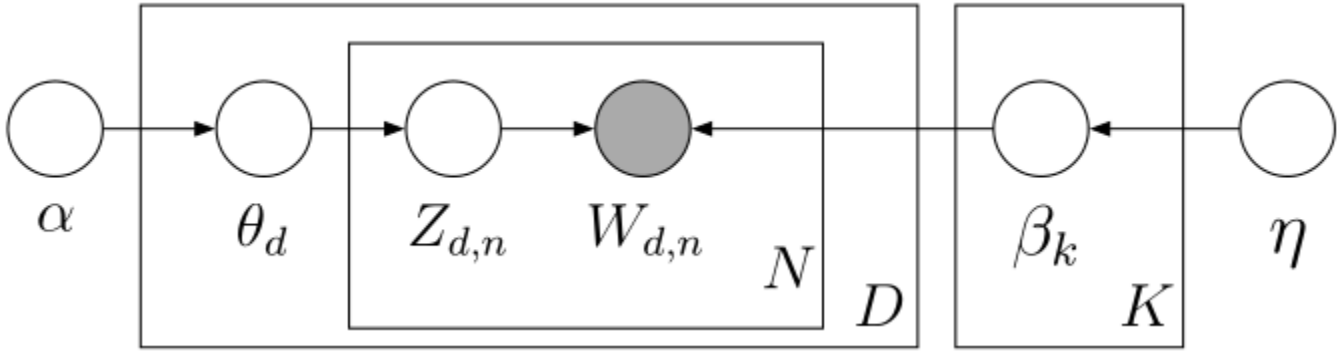
- Jedes Wort wird aus einem der Themen erzeugt:
  - Für jedes Dokument $d \in \{1,\ldots,D\}$ und jedes Wort $n \in \{1,\ldots,N_d\}$ ist $Z_{d,n} \in \{1,\ldots,K\}$ eine *1*-dimensionale Zufallsvariable.
  - $Z_{d,n}$ legt das Thema an der Stelle n in Dokument d fest.
  - $Z_{d,n} \mid \theta_d \sim \text{Mult}(\theta_d)$ ist multinomialverteilt, wobei θ_d die Einzelwahrscheinlichkeiten spezifiziert.
  - Für jedes Dokument $d \in \{1,\ldots,D\}$ und jedes Wort $n \in \{1,\ldots,N_d\}$ ist $W_{d,n} \in \{1,\ldots,V\}$ eine *1*-dimensionale Zufallsvariable.
  - $W_{d,n}$ legt das Wort an der Stelle n in Dokument d fest, gegeben das Thema dieser Stelle.
  - $W_{d,n} \mid Z_{d,n}, \beta \sim \text{Mult}(\beta_{Z_{d,n}})$ ist multinomialverteilt, wobei $\beta_{Z_{d,n}}$ die Einzelwahrscheinlichkeiten spezifiziert.

Schreibweise für:
$P(W_{d,n} \mid Z_{d,n}, \beta)$ entspricht der Wahrscheinlichkeitsfunktion einer Multinomialverteilung mit Parameter $\beta_{Z_{d,n}}$

Latente Dirichlet-Allokation



Latente Dirichlet-Allokation



- Modell spezifiziert bedingte Unabhängigkeiten:

$$P(\theta, Z, W, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \times$$

$$\prod_{d=1}^{D} \prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n} \mid Z_{d,n}, \beta)$$

- Hyperparameter α und η sind fest

Latente Dirichlet-Allokation

$$P(\theta, Z, W, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \times \prod_{d=1}^{D} \prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n} \mid Z_{d,n}, \beta)$$

- Modellierung der einzelnen Verteilungen:
  - $Z_{d,n} \mid \theta_d \sim \text{Mult}(\theta_d)$ ist multinomialverteilt.
  - $W_{d,n} \mid Z_{d,n}, \beta \sim \text{Mult}(\beta_{Z_{d,n}})$ ist multinomialverteilt.
  - Wie sollte man $P(\beta_k \mid \eta)$ und $P(\theta_d \mid \alpha)$ wählen?
    - Konjugierte Verteilung zu Mutlinomialverteilung

Wiederholung: Dirichlet-Verteilung

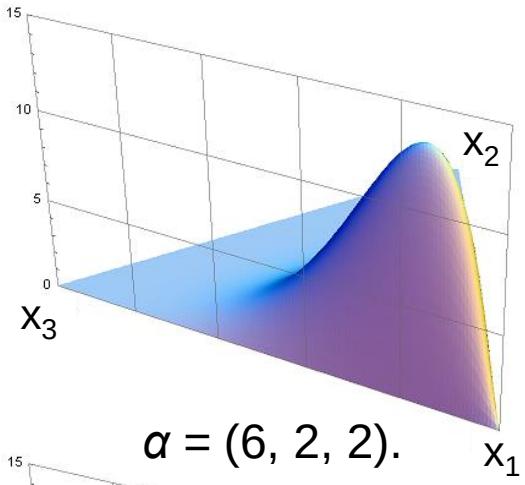
- Verteilung über Vektoren $\theta = (\theta_1, \ldots, \theta_D)$, mit:
  - Alle Einträge $\theta_d > 0$ sind positiv,
  - Die Summe der Einträge ist $\sum_{d=1}^{D} \theta_d = 1$.
- Verallgemeinerung der Beta-Verteilung auf mehr als 2 Dimensionen
- Parametrisiert durch Vektor $\alpha = (\alpha_1, \ldots, \alpha_D)$ mit $\alpha_d > 0$
- Dichtefunktion:

$$P(\theta \mid \alpha_1, \ldots, \alpha_D) = \frac{\prod_{i=1}^{D} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{D} \alpha_i\right)} \prod_{i=1}^{D} \theta_i^{\alpha_i - 1}$$

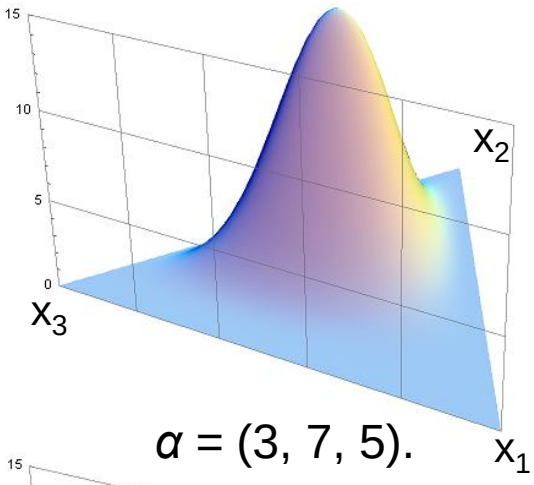
Verallgemeinerte Beta-Funktion

Gamma-Funktion

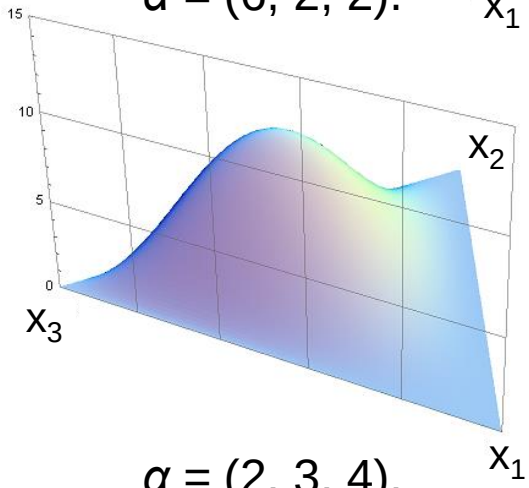
Wiederholung: Dirichlet-Verteilung



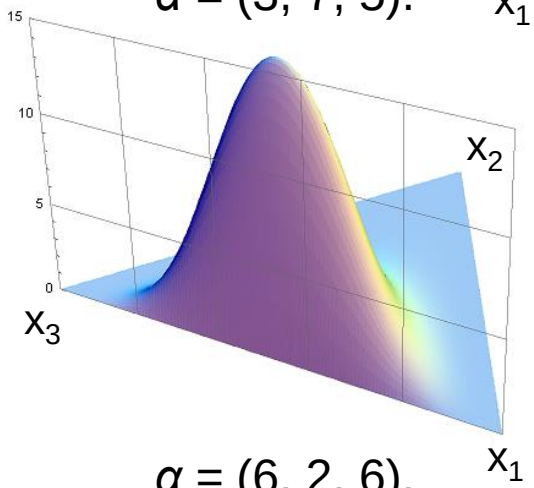
$\alpha = (6, 2, 2)$.



$\alpha = (3, 7, 5)$.



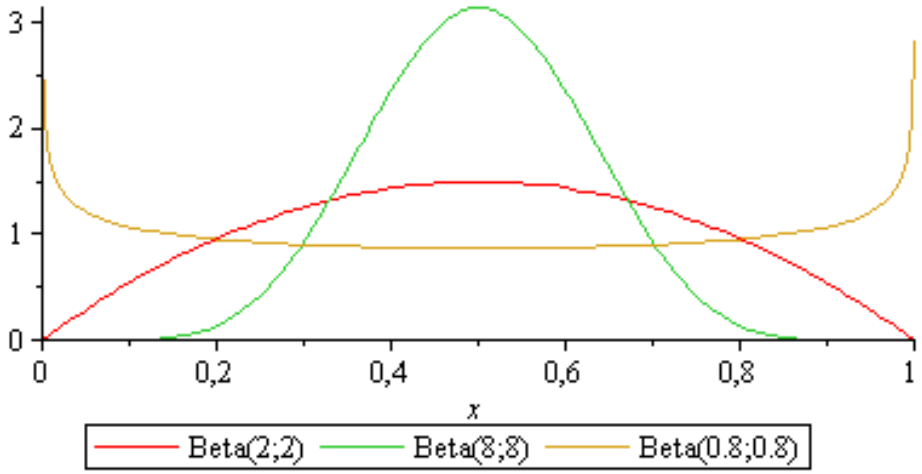
$\alpha = (2, 3, 4)$.



$\alpha = (6, 2, 6)$.

Wiederholung: Dirichlet-Verteilung

- Je größer die Summe der Alphas, desto „spitzer“ ist die Verteilung
  - Man erhält schwach variierende Vektoren
- Je kleiner die Summe der Alphas, desto mehr Wahrscheinlichkeitsmasse konzentriert sich auf die Ränder und Ecken
  - Man erhält sparse Vektoren (meiste Komponenten 0)



Wiederholung: Dirichlet-Verteilung

- Dirichlet-Verteilung ist der konjugierte Prior der Multinomialverteilung
  - Posterior hat dieselbe Form wie der Prior
- Bei LDA: normalerweise austauschbarer Dirichletprior
  - alle Komponenten des Parametervektors identisch
  - effektiv nur ein Parameter

$$P(\theta \mid \alpha) = \frac{\Gamma(\alpha)^D}{\Gamma(D \cdot \alpha)} \prod_{i=1}^{D} \theta_i^{\alpha - 1}$$

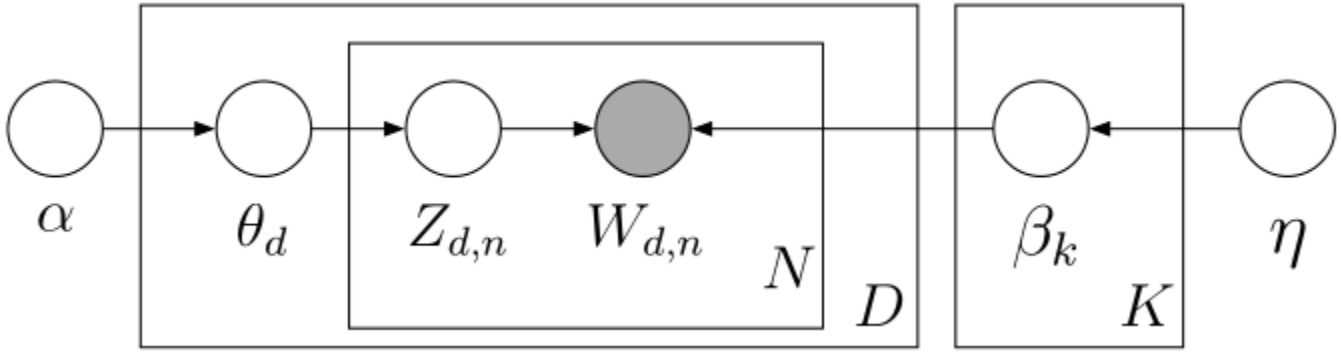
Gamma-Funktion

Latente Dirichlet-Allokation

$$P(\theta, Z, W, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \times \prod_{d=1}^{D} \prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n} \mid Z_{d,n}, \beta)$$

- Modellierung der einzelnen Verteilungen:
  - $Z_{d,n} \mid \theta_d \sim \text{Mult}(\theta_d)$ ist multinomalverteilt.
  - $W_{d,n} \mid Z_{d,n}, \beta \sim \text{Mult}(\beta_{Z_{d,n}})$ ist multinomalverteilt.
  - $\beta_k \mid \eta \sim \text{Dir}(\eta)$ ist Dirichlet-verteilt.
  - $\theta_d \mid \alpha \sim \text{Dir}(\alpha)$ ist Dirichlet-verteilt.
    - Wichtig: $\alpha < 1$, damit nur wenige Themen pro Dokument vergeben werden

LDA: Teilprobleme



- Aus gegebener Dokumentensammlung, inferiere Posterior-Verteilung von
  - Themenzuweisungen für jedes Wort $z_{d,n}$
  - Themenanteile für jedes Dokument θ_d
  - Verteilung über Vokabular für jedes Thema β_k
- Benutze Erwartungswerte des Posteriors für verschiedene Anwendungen

Posterior-Inferenz

- Berechnung der Posterior-Verteilung zu schwierig:

$$P(\theta, Z, \beta \mid W, \alpha, \eta) = \frac{P(\theta, Z, W, \beta \mid \alpha, \eta)}{\int_{\theta}\int_{\beta}\sum_{Z} P(\theta, Z, W, \beta \mid \alpha, \eta) d\beta d\theta}$$

- Daher: Approximative Posterior-Inferenz
- Mehrere Möglichkeiten:
  - Mean-field-variational-Methoden
  - Expectation propagation
  - Collapsed variational inference
  - Gibbs-Sampling
    - einfachstes Verfahren

Gibbs-Sampling

- Echte Posteriorverteilung $P(\theta, Z, \beta | W, \alpha, \eta)$ zu schwierig
- Gibbs-Sampling produziert I Samples aus der echten Verteilung

$$\left(\theta^1, Z^1, \beta^1\right), \ldots, \left(\theta^I, Z^I, \beta^I\right) \sim P\left(\theta, Z, \beta \,|\, W, \alpha, \eta\right)$$

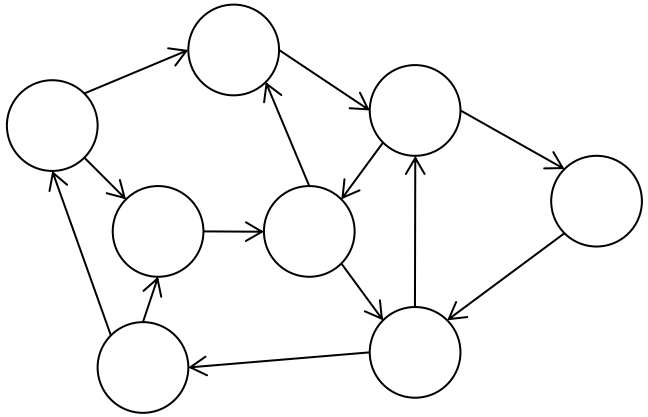
- Idee: Posterior für einzelne Zufallsvariablen ist leicht zu berechnen, z.B.

$$P(z_{7,15} | \theta, Z \setminus \{z_{7,15}\}, \beta, W, \alpha, \eta)$$

  - nacheinander jede Zufallsvariable neu aus ihrem Posterior gegeben alle anderen Variablen ziehen

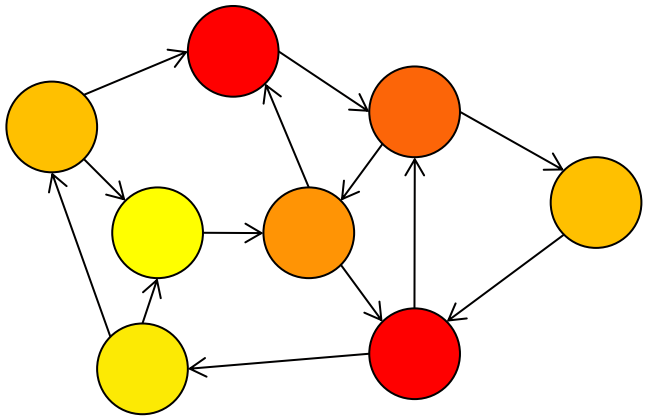
Gibbs-Sampling

- Ein MCMC (Markov-Chain-Monte-Carlo) - Algorithmus
- Neu Ziehen von einzelnen Variablen entspricht Zustandsübergang
- Zustand = komplette Belegung von allen Variablen
- Iteratives neu ziehen ist Random Walk auf dem Zustandsgraphen

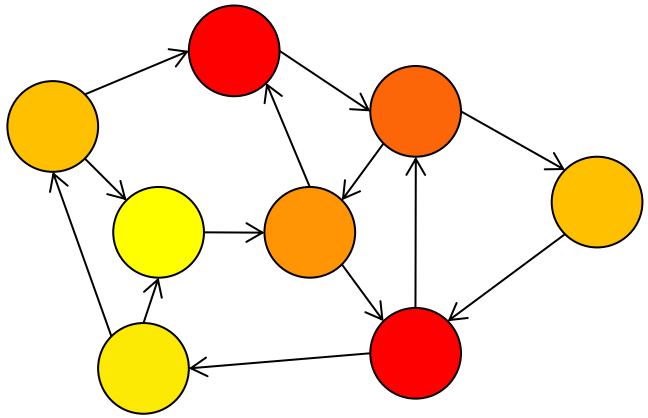


Gibbs-Sampling

- Häufigkeiten des Besuchs von Zuständen entspricht Verteilung über die Belegungen der Zufallsvariablen



Gibbs-Sampling



- Theorem: Wenn man sich lange genug auf dem Zustandsgraphen bewegt, und die Zustandsübergänge aus den Einzel-Posteriors zieht, konvergieren die Besuchshäufigkeiten zur Gesamt-Posterior-Verteilung.
  - Bemerkung: Der Startpunkt ist spielt keine Rolle.

Gibbs-Sampling

- Um Samples $\left(\theta^1, Z^1, \beta^1\right), \ldots, \left(\theta^I, Z^I, \beta^I\right) \sim P\left(\theta, Z, \beta \mid W, \alpha, \eta\right)$ aus dem Gesamtposterior zu erhalten:
  - mit beliebigem Startwert beginnen, und samplen, bis die Verteilung konvergiert
  - zwischen dem Entnehmen einzelner Samples viele Sampleschritte durchführen, damit die Samples voneinander unabhängig sind

Gibbs-Sampling: Algorithmus

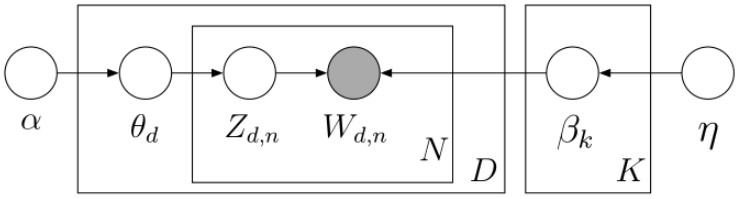
- Für i von 1 bis I
  - Für j von 1 bis 1000 (z.B.)
    - Für alle Themen k:
      - Ziehe $\beta_k \sim P(\beta_k|\theta, Z, \beta_{-k}, W, \alpha, \eta)$
    - Für alle Dokumente d:
      - Ziehe $\theta_d \sim P(\theta_d|\theta_{-d}, Z, \beta, W, \alpha, \eta)$
      - Für alle Wörter n in Dokument d:
        - Ziehe $z_{d,n} \sim P(z_{d,n}|\theta, Z_{-d,n}, \beta, W, \alpha, \eta)$
  - Gib Posterior-Sample (θ^i, Z^i, β^i) aus.

Collapsed Gibbs-Sampling

- Bei normalem Gibbs-Sampling: sehr viele Iterationen notwendig
- Verbesserung: nur Z samplen, θ und β rausintegrieren
  - Idee dahinter: kleinerer Zustandsgraph, da Zustand nur noch aus Z besteht
  - Statt $z_{d,n} \sim P\left(z_{d,n} \mid \theta, Z_{-d,n}, \beta, W, \alpha, \eta\right)$:

$$z_{d,n} \sim P\left(z_{d,n} \mid Z_{-d,n}, W, \alpha, \eta\right)$$
$$= \int_{\theta}\int_{\beta} P\left(z_{d,n} \mid \theta, Z_{-d,n}, \beta, W, \alpha, \eta\right) P\left(\theta, \beta \mid Z_{-d,n}, W, \alpha, \eta\right) d\beta d\theta$$

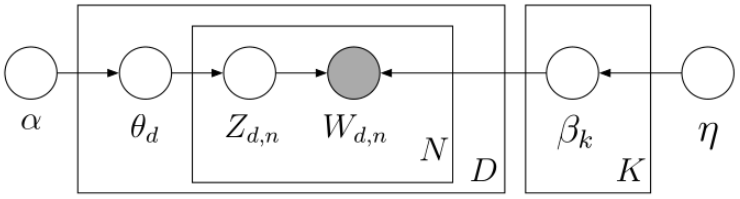
Collapsed Gibbs-Sampling



$$P(\theta, Z, W, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{d=1}^{D} \prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n} \mid Z_{d,n}, \beta)$$

$P(z_{d,n} \mid Z_{-d,n}, W, \alpha, \eta)$	Produktregel
$= P(z_{d,n}, Z_{-d,n}, W \mid \alpha, \eta) / P(Z_{-d,n}, W \mid \alpha, \eta)$	Zähler
$\propto P(z_{d,n}, Z_{-d,n}, W \mid \alpha, \eta)$	Produktregel
$= P(w_{d,n} \mid z_{d,n}, Z_{-d,n}, W_{-d,n}, \alpha, \eta) P(z_{d,n}, Z_{-d,n}, W_{-d,n} \mid \alpha, \eta)$	Produktregel
$= P(w_{d,n} \mid z_{d,n}, Z_{-d,n}, W_{-d,n}, \alpha, \eta) P(z_{d,n} \mid Z_{-d,n}, W_{-d,n}, \alpha, \eta) \times P(Z_{-d,n}, W_{-d,n} \mid \alpha, \eta)$	Letzten Faktor ignorieren
$\propto P(w_{d,n} \mid z_{d,n}, Z_{-d,n}, W_{-d,n}, \alpha, \eta) P(z_{d,n} \mid Z_{-d,n}, W_{-d,n}, \alpha, \eta)$	Bedingte Unabhängigkeit
$= P(w_{d,n} \mid z_{d,n}, Z_{-d,n}, W_{-d,n}, \eta) P(z_{d,n} \mid Z_{-d,n}, \alpha)$	

Collapsed Gibbs-Sampling



$$P(\theta, Z, W, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{d=1}^{D} \prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n} \mid Z_{d,n}, \beta)$$

- 1. Term („Likelihood“):

$$P(w_{d,n} = w \mid z_{d,n} = z, Z_{-d,n}, W_{-d,n}, \eta)$$

Randverteilung, Produktregel, Unabhängigkeit

$$= \int_{\beta_z} P(w_{d,n} = w \mid z_{d,n} = z, \beta_z) P(\beta_z \mid W_{-d,n}, Z_{-d,n}, \eta) d\beta_z$$

Definitionen

$$= \int_{\beta_z} \beta_{z,w} \mathrm{Dir}(\beta_z \mid \eta + \mathrm{counts}(z, \cdot)) d\beta_z$$

Erwartungswert Dirichlet

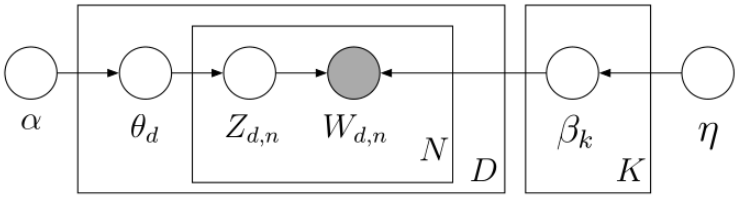
Zähler, wie oft Wort w Thema z zugewiesen ist (ohne $w_{d,n}$)

$$= \frac{\mathrm{counts}(z, w) + \eta}{\mathrm{counts}(z) + V\eta}$$

Größe des Vokabulars

Zähler, wie viele Wörter Thema z zugewiesen sind (ohne $w_{d,n}$)

Collapsed Gibbs-Sampling



$$P(\theta, Z, W, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{d=1}^{D} \prod_{n=1}^{N} P(Z_{d,n} \mid \theta_d) P(W_{d,n} \mid Z_{d,n}, \beta)$$

- 2. Term („Prior“):

$$P(z_{d,n} = z | Z_{-d,n}, \alpha)$$

Randverteilung, Produktregel, Unabhängigkeit

$$= \int_{\theta_d} P(z_{d,n} = z | \theta_d) P(\theta_d | Z_{-d,n}, \alpha) d\theta_d$$

Definitionen

$$= \int_{\theta_d} \theta_{d,z} P_{Dirichlet}(\theta_d | \alpha + counts(d, \cdot)) d\theta_d$$

Erwartungswert Dirichlet

Zähler, wie viele Wörter in d Thema z zugewiesen sind (ohne $w_{d,n}$)

$$= \frac{count(d, z) + \alpha}{count(d) + K\alpha}$$

Anzahl der Themen

Anzahl der Wörter in d (ohne $w_{d,n}$)

Collapsed Gibbs-Sampling

$$P(z_{d,n} = z | Z_{-d,n}, W, \alpha, \eta) \propto \frac{count(z,w) + \eta}{count(z) + V\eta} \times \frac{count(d,z) + \alpha}{count(d) + K\alpha}$$

- Effizient implementierbar
- Man muss sich immer nur die 4 verschiedenen Typen von Zählern merken
- Bei Bedarf lassen sich zusätzlich Samples für θ und β generieren
  - Einfach aus Posterior $P(\beta_z | W, Z, \eta)$ oder $P(\theta_d | Z, \alpha)$ ziehen
- Kleinerer Zustandsgraph $\rightarrow$ schnelleres Einpendeln auf richtige Verteilung

Anwendung (Eingangsbeispiel)

Chance and Statistical Significance in Protein and DNA Sequence Analysis

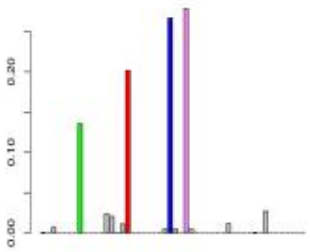
Samuel Karlin and Volker Brendel

$E[\beta_z \mid W]$

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



$E[\theta_d \mid W]$

Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal genera compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

$E[z_{d,n} = z \mid W]$

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Ähnlichkeit der Themen (nächste Folie)

Anwendung (Ähnlichkeit von Dokumenten)

- Berechnung der Ähnlichkeit zweier Dokumente als inneres Produkt der Themenverteilungen

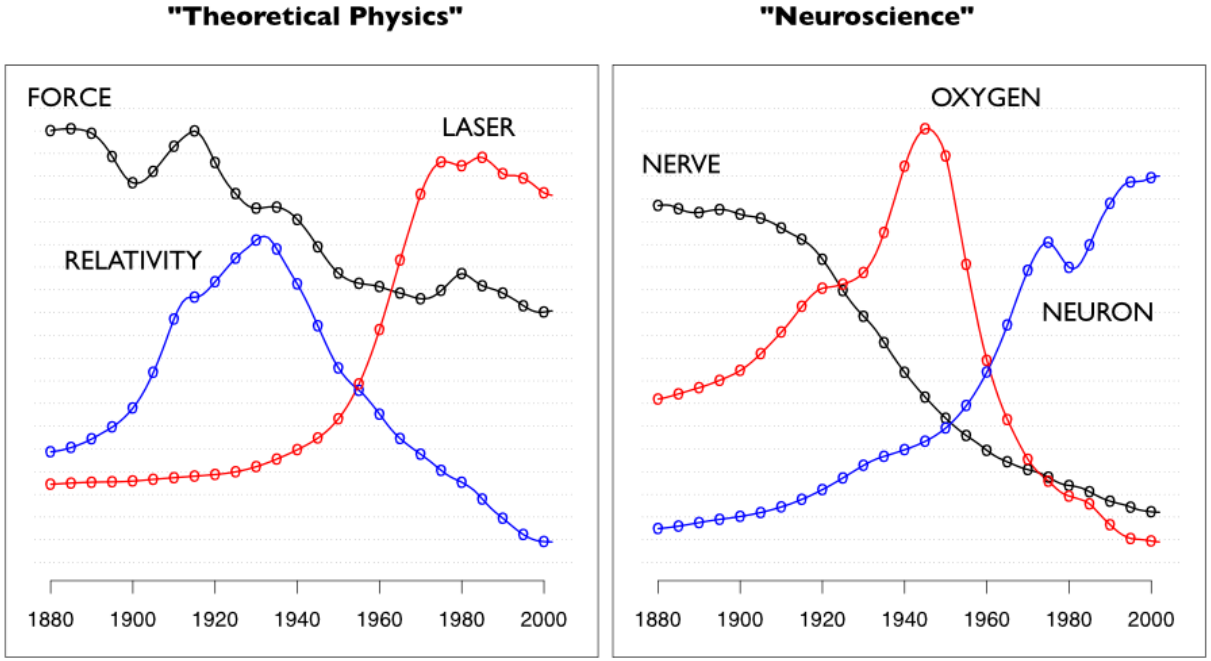
Bayesian Model Averaging

$$sim(d_j, d_l) = \int_{\theta,Z,\beta} \langle \theta_j, \theta_l \rangle \, P(\theta, Z, \beta | W, \alpha, \eta) d\theta dZ d\beta$$

$$\approx \frac{1}{I} \sum_{i=1}^{I} \langle \theta_j^i, \theta_l^i \rangle$$

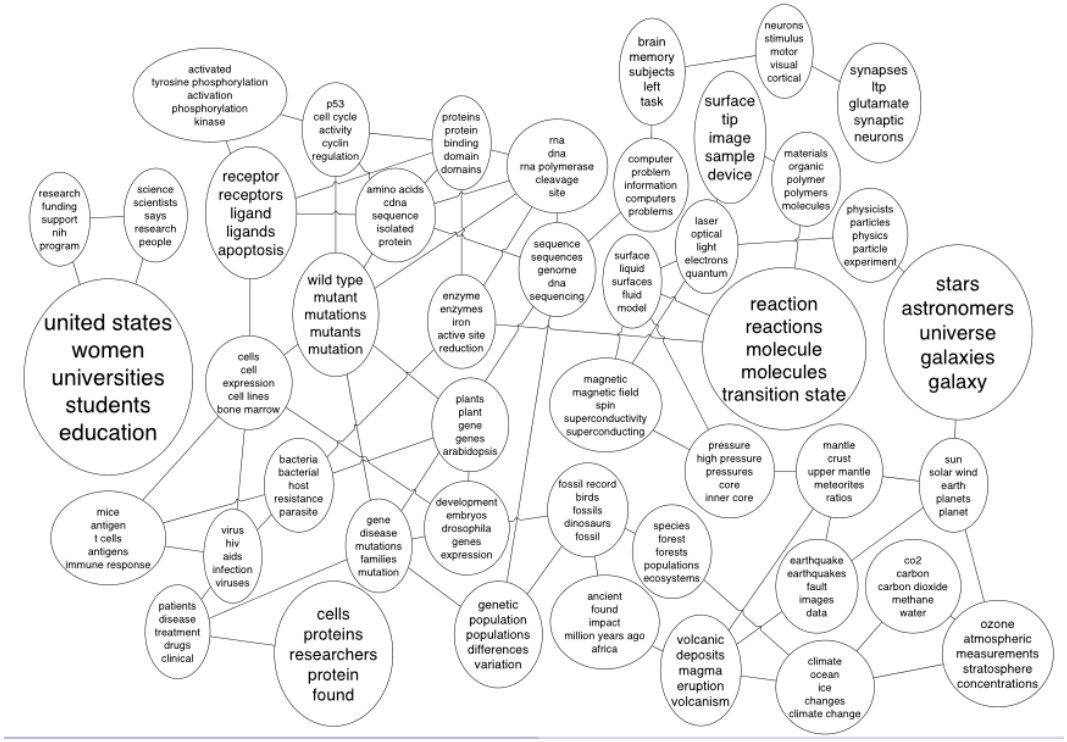
Ersetzen des Integrals über die Posterior-Verteilung durch Summe über Gibbs-Samples aus dem Posterior

Erweiterung (Entwicklung von Themen über die Zeit)



- Variablen $\beta_{z,t}$ verändern sich über Zeit $t=1,...,$

Erweiterung (Zusammenhänge zwischen Themen)



- Darstellung von:
  - den 5 häufigsten Wörtern pro Thema,
  - der Häufigkeit des Themas im Corpus (Schriftgröße)
  - Covarianz der Themen (Graph)

Fragen?

Acknowledgements

- Folien basieren teilweise auf Tutorial von David Blei, Machine Learning Summer School 2009