Data Mining and Matrices

12 – Probabilistic Matrix Factorization

Rainer Gemulla, Pauli Miettinen

Jul 18, 2013

Why probabilistic?

- Until now, we factored the data $\mathbf{D}$ in terms of factor matrices $\mathbf{L}$ and $\mathbf{R}$ such that

$$\mathbf{D} \approx \mathbf{LR},$$

  subject to certain constraints

- We (somewhat) skimmed over questions like
  - Which assumptions underly these factorizations?
  - What is the meaning of parameters? How can we pick them?
  - How can we quantify the uncertainty in the results?
  - How can we deal with new rows and new columns?
  - How can we add background knowledge to the factorization?

- Bayesian treatments of matrix factorization models help answer these questions

Outline

1. Background: Bayesian Networks
2. Probabilistic Matrix Factorization
3. Latent Dirichlet Allocation
4. Summary

What do probabilities mean?

- Multiple interpretations of probability
- **Frequentist interpretation**
  - Probability of an event = relative frequency when repeated often
  - Coin, n trials, n_{H} observed heads

$$\lim_{n\to\infty} \frac{n_{\mathrm{H}}}{n} = \frac{1}{2} \implies \mathbb{P}(\mathrm{H}) = \frac{1}{2}$$

- **Bayesian interpretation**
  - Probability of an event = degree of belief that event holds
  - Reasoning with "background knowledge" and "data"
  - Prior belief + model + data $\rightarrow$ posterior belief
    - ★ Model parameter: θ = true "probability" of heads
    - ★ Prior belief: $\mathbb{P}(\theta)$
    - ★ Likelihood (model): $\mathbb{P}(n_{\mathrm{H}}, n \mid \theta)$
    - ★ Posterior belief: $\mathbb{P}(\theta \mid n_{\mathrm{H}}, n)$
    - ★ Bayes theorem: $\mathbb{P}(\theta \mid n_{\mathrm{H}}, n) \propto \mathbb{P}(n_{\mathrm{H}}, n \mid \theta)\,\mathbb{P}(\theta)$

Bayesian methods make use of a probabilistic model (priors + likelihood) and the data to infer the posterior distribution of unknown variables.

Probabilistic models

- Suppose you want to diagnose diseases of a patient
- Multiple interrelated aspects may relate to the reasoning task
  - Possible diseases, hundreds of symptoms and diagnostic tests, personal characteristics, ...

1. Characterize data by a set of **random variables**
   - Flu (yes / no)
   - Hayfever (yes / no)
   - Season (Spring / Sommer / Autumn / Winter)
   - Congestion (yes / no)
   - MusclePain (yes / no)

   $\rightarrow$ Variables and their domain are important design decision
2. Model dependencies by a **joint distribution**
   - Diseases, season, and symptoms are correlated
   - Probabilistic models construct joint probability space
     $\rightarrow 2 \cdot 2 \cdot 4 \cdot 2 \cdot 2$ outcomes (64 values, 63 non-redundant)
   - Given joint probability space, interesting questions can be answered

$$\mathbb{P}(\,\text{Flu} \mid \text{Season}{=}\text{Spring}, \text{Congestion}, \neg\text{MusclePain}\,)$$

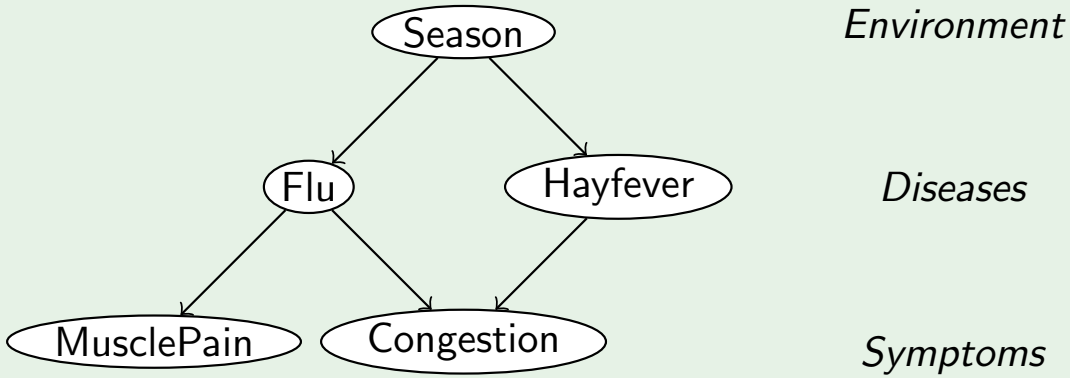
Specifying a joint distribution is infeasible in general!

Bayesian networks are ...

- A graph-based representation of **direct probabilistic interactions**
- A break-down of high-dimensional distributions into smaller **factors** (here: 63 vs. 17 non-redundant parameters)
- A compact representation of (cond.) **independence assumptions**

Example (directed graphical model)

Graph representation



Factorization

$$\mathbb{P}(S, F, H, M, C)$$
$$= \mathbb{P}(S)\,\mathbb{P}(F \mid S)\,\mathbb{P}(H \mid S)\,\mathbb{P}(C \mid F, H)\,\mathbb{P}(M \mid F)$$

Independencies

$$(F \perp H \mid S),\ (C \perp S, M \mid F, H),\ \ldots$$

Independence (events)

Definition

Two events A and B are called **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

If $\mathbb{P}(B) > 0$, implies that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

Example (fair die)

Two independent events:

- Die shows an even number: $A = \{2, 4, 6\}$
- Die shows at most 4: $B = \{1, 2, 3, 4\}$:
- $\mathbb{P}(A \cap B) = \mathbb{P}(\{2, 4\}) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = \mathbb{P}(A)\,\mathbb{P}(B)$

Not independent:

- Die shows at most 3: $B = \{1, 2, 3\}$
- $\mathbb{P}(A \cap B) = \mathbb{P}(\{2\}) = \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B)$

Conditional independence (events)

Definition

Let A, B, C be events with $\mathbb{P}(C) > 0$. A and B are **conditionally independent given** C if $\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$.

Example

Not independent:

- Die shows an even number: $A = \{2, 4, 6\}$
- Die shows at most 3: $B = \{1, 2, 3\}$
- $\mathbb{P}(A \cap B) = \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B)$
  $\rightarrow$ A and B are not independent

Conditionally independent:

- Die does not show multiple of 3: $C = \{1, 2, 4, 5\}$
- $\mathbb{P}(A \cap B \mid C) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$
  $\rightarrow$ A and B are conditionally independent given C

Shortcut notation

Let X and Y be discrete random variables with domain $\mathrm{Dom}(X)$ and $\mathrm{Dom}(Y)$. Let $x \in \mathrm{Dom}(X)$ and $y \in \mathrm{Dom}(Y)$.

Expression	Shortcut notation
$\mathbb{P}(X = x)$	$\mathbb{P}(x)$
$\mathbb{P}(X = x \mid Y = y)$	$\mathbb{P}(x \mid y)$
$\forall x.\, \mathbb{P}(X = x) = f(x)$	$\mathbb{P}(X) = f(X)$
$\forall x. \forall y.\, \mathbb{P}(X = x \mid Y = y) = f(x, y)$	$\mathbb{P}(X \mid Y) = f(X, Y)$

- $\mathbb{P}(X)$ and $\mathbb{P}(X \mid Y)$ are entire probability distributions
- Can be thought of as functions from $\mathrm{Dom}(X) \to [0, 1]$ or $(\mathrm{Dom}(X), \mathrm{Dom}(Y)) \to [0, 1]$, respectively
- $f_y(X) = \mathbb{P}(X \mid y)$ is often referred to as **conditional probability distribution** (CPD)
- For finite discrete variables, may be represented as a table (CPT)

Important properties

Let A, B be events, and let X, Y be discrete random variables.

Theorem

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \textit{(inclusion-exclusion)}$$

$$\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$$

$$\textit{If } B \supseteq A, \quad \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$$

$$\mathbb{P}(X) = \sum_{y} \mathbb{P}(X, Y = y) \quad \textit{(sum rule)}$$

$$\mathbb{P}(X, Y) = \mathbb{P}(Y \mid X)\,\mathbb{P}(X) \quad \textit{(product rule)}$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \quad \textit{(Bayes theorem)}$$

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b \quad \textit{(linearity of expectation)}$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X] \quad \textit{(law of total expectation)}$$

Conditional independence (random variables)

Definition

Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be sets of discrete random variables. $\mathcal{X}$ and $\mathcal{Y}$ are said to be **conditionally independent** given $\mathcal{Z}$ if and only if

$$\mathbb{P}(\mathcal{X}, \mathcal{Y} \mid \mathcal{Z}) = \mathbb{P}(\mathcal{X} \mid \mathcal{Z})\,\mathbb{P}(\mathcal{Y} \mid \mathcal{Z}).$$

We write $(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z})$ for this conditional independence statement. If $\mathcal{Z} = \emptyset$, we write $(\mathcal{X} \perp \mathcal{Y})$ for **marginal independence**.

Example

- Throw a fair coin: $Z = 1$ if head, else $Z = 0$
- Throw again: $X = Z$ if head, else $X = 0$
- Throw again: $Y = Z$ if head, else $Y = 0$
- $\mathbb{P}(X = 0, Y = 0 \mid Z = 0) = 1 = \mathbb{P}(X = 0 \mid Z = 0)\,\mathbb{P}(Y = 0 \mid Z = 0)$
- $\mathbb{P}(x, y \mid Z = 1) = 1/4 = \mathbb{P}(x \mid Z = 1)\,\mathbb{P}(y \mid Z = 1)$
- Thus $(X \perp Y \mid Z)$, but note $(X \not\perp Y)$

Properties of conditional independence

Theorem

In general, $(\mathcal{X} \perp \mathcal{Y})$ *does not imply nor is implied by* $(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z})$.

The following relationships hold:

$$
\begin{aligned}
(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}) &\iff (\mathcal{Y} \perp \mathcal{X} \mid \mathcal{Z}) && (\textit{symmetry}) \\
(\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) &\implies (\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}) && (\textit{decomposition}) \\
(\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) &\implies (\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}, \mathcal{W}) && (\textit{weak union}) \\
(\mathcal{X} \perp \mathcal{W} \mid \mathcal{Z}, \mathcal{Y}) \wedge (\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}) &\implies (\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) && (\textit{contraction})
\end{aligned}
$$

For positive distributions and mutally disjoint sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}$*:*

$$(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}, \mathcal{W}) \wedge (\mathcal{X} \perp \mathcal{W} \mid \mathcal{Z}, \mathcal{Y}) \implies (\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) \quad (\textit{intersection})$$

Bayesian network structure

Definition

A **Bayesian network structure** is a directed acyclic graph $\mathcal{G}$ whose nodes represent random variables $\mathcal{X} = \{ X_1, \ldots, X_n \}$. Let

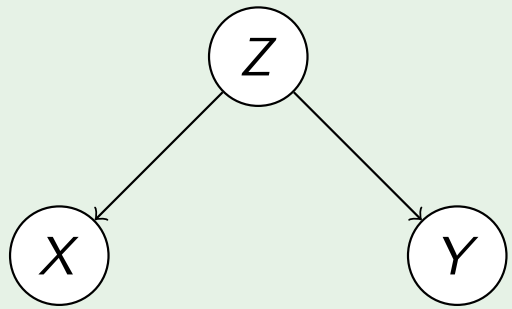
- Pa_{X_i} = set of parents of X_i in $\mathcal{G}$,
- $\text{NonDescendants}_{X_i}$ = set of variables that are not descendants of X_i.

$\mathcal{G}$ encodes the following **local independence assumptions**:

$$(X_i \perp \text{NonDescendants}_{X_i} \mid \text{Pa}_{X_i}) \qquad \text{for all } X_i.$$

Example

- $\text{Pa}_Z = \emptyset$, $\text{Pa}_X = \text{Pa}_Y = \{ Z \}$
- $\text{NonDescendants}_X = \{ Y, Z \}$
- $\text{NonDescendants}_Y = \{ X, Z \}$
- $\text{NonDescendants}_Z = \emptyset$
- $(X \perp Y, Z \mid Z) \xrightarrow{\text{decomposition}} (X \perp Y \mid Z)$



Factorization

Definition

A distribution $\mathbb{P}$ over $X_1, \ldots, X_n$ **factorizes over** $\mathscr{G}$ if it can be written as

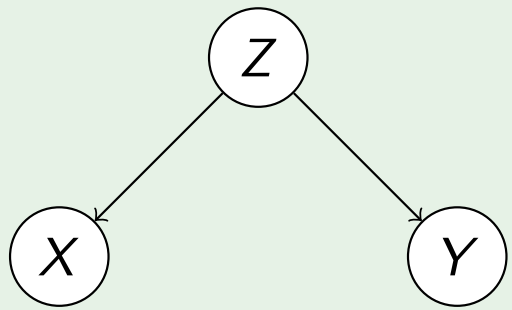
$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \mathrm{Pa}_{X_i}). \qquad (\textit{chain rule})$$

Theorem

$\mathbb{P}$ factorizes over $\mathscr{G}$ if and only if $\mathbb{P}$ satisfies the local independence assumptions of $\mathscr{G}$.

Example

- $\mathbb{P}(X, Y, Z) = \mathbb{P}(Z)\,\mathbb{P}(X \mid Z)\,\mathbb{P}(Y \mid Z)$
- $(X \perp Y \mid Z)$
- Holds for 3-coin example from slide 11
- Holds for 3 independent coin throws
- Doesn't hold: throw Z; throw again and set $X = Y = Z$ if head, else 0



Bayesian network

Definition

A **Bayesian network** is a pair $(\mathscr{G}, \mathbb{P})$, where $\mathbb{P}$ factorizes of $\mathscr{G}$ and $\mathbb{P}$ is given as a set of **conditional probability distributions** (CPDs)

$$\mathbb{P}(X_i \mid \mathrm{Pa}_{X_i}) \qquad \text{for all } X_i.$$

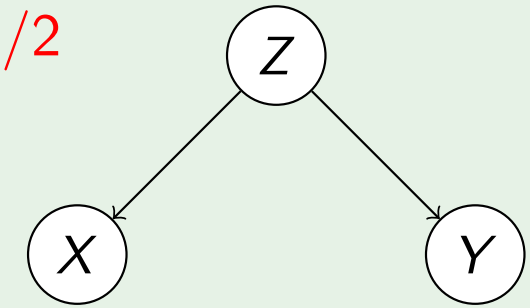
Example

z	$\mathbb{P}(z)$
0	1/2
1	1/2

x	z	$\mathbb{P}(x \mid z)$
0	0	1
1	0	0
0	1	1/2
1	1	1/2

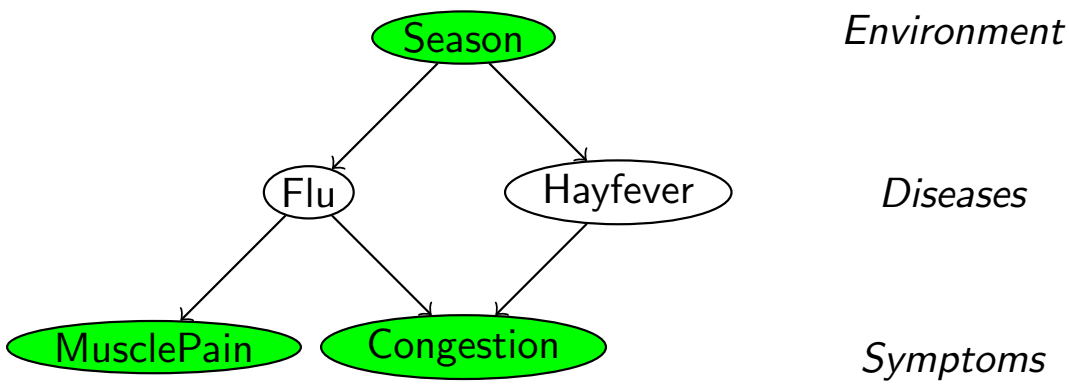
y	z	$\mathbb{P}(y \mid z)$
0	0	1
1	0	0
0	1	1/2
1	1	1/2

- redundant
- CPDs: 5 non-redundant parameters
- Full distribution: 7 non-redundant parameters



Generative models

- Bayesian networks describe how to generate data: **forward sampling**
  1. Pick S: Which season is it? ($\mathbb{P}(S)$)
  2. Pick F: Does the patient have flu? ($\mathbb{P}(F \mid S)$)
  3. Pick H: Does the patient have hayfever? ($\mathbb{P}(H \mid S)$)
  4. Pick M: Does the patient have muscle pain? ($\mathbb{P}(M \mid F)$)
  5. Pick C: Does the patient have congestion? ($\mathbb{P}(C \mid F, H)$)
- Hence are often called **generative models**
  - Encode modeling assumptions (independencies, form of distributions)
- In practice, we do not want to generate data
  - Some variables are observed
  - Goal is to infer properties of the other variables



Querying a distribution (1)

Consider a joint distribution on a set of variables $\mathcal{X}$

- Let $\mathcal{E} \subseteq \mathcal{X}$ be a set of **evidence variables** that takes values e
- Let $\mathcal{W} = \mathcal{X} \setminus \mathcal{E}$ be the set of **latent variables**
- Let $\mathcal{Y} \subseteq \mathcal{W}$ be a set of **query variables**
- Let $\mathcal{Z} = \mathcal{W} \setminus \mathcal{Y}$ be the set of **non-query variables**

Example

- $\mathcal{X} = \{\,\text{Season}, \text{Congestion}, \text{MusclePain}, \text{Flu}, \text{Hayfever}\,\}$
- $\mathcal{E} = \{\,\text{Season}, \text{Congestion}, \text{MusclePain}\,\}$
- $e = \{\,\text{Season: Spring}, \text{Congestion: Yes}, \text{MusclePain: No}\,\}$
- $\mathcal{W} = \{\,\text{Flu}, \text{Hayfever}\,\}$
- $\mathcal{Y} = \{\,\text{Flu}\,\}$
- $\mathcal{Z} = \{\,\text{Hayfever}\,\}$

Querying a distribution (2)

1. **Conditional probability query**
   - Compute the posterior distribution of the query variables
     $\mathbb{P}(\mathcal{Y} \mid e)$
2. **MAP query**
   - Compute the most likely value of the latent variables
     $\text{MAP}(\mathcal{W} \mid e) = \text{argmax}_w \mathbb{P}(w \mid e) = \text{argmax}_w \mathbb{P}(w, e)$
3. **Marginal MAP query**
   - Compute the most likely value of the query variables
     $\text{MAP}(\mathcal{Y} \mid e) = \text{argmax}_y \mathbb{P}(y \mid e) = \text{argmax}_y \sum_z \mathbb{P}(y, z, e)$

Example

$\mathbb{P}(\mathcal{W} \mid e)$	Flu	¬Flu
Hayfever	5%	35%
¬Hayfever	40%	20%

1. $\mathbb{P}(\text{Flu} \mid \text{Spring}, \text{Congestion}, \neg\text{MusclePain}) \rightarrow$ Yes (45%), No (55%)
2. $\text{MAP}(\text{Flu}, \text{Hayfever} \mid \text{Spring}, \text{Congestion}, \neg\text{MusclePain}) \rightarrow$ Only flu
3. $\text{MAP}(\text{Flu} \mid \text{Spring}, \text{Congestion}, \neg\text{MusclePain}) \rightarrow$ No flu (!)

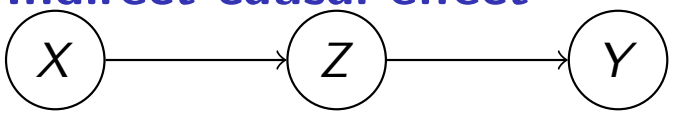
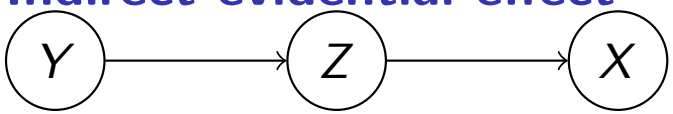
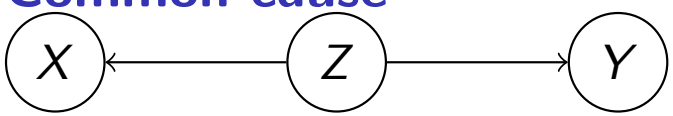
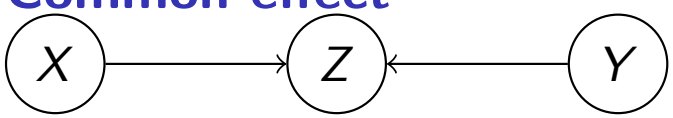
Probabilistic inference

- Probabilistic inference = compute (properties of) posterior $\mathbb{P}(\mathcal{Y} \mid e)$
- Example: use forward sampling (naive)
  1. Sample from the BN
  2. Drop sample if does not agree with evidence
  3. Repeat until sufficiently many samples have been retained
  4. Investigate the values of the latent variables in these samples

  $\rightarrow$ This usually does not scale (unless evidence at "roots" only)
- Many methods (not discussed here)
  - Variable elimination
  - Message passing methods
  - Markov-Chain Monte Carlo methods
  - Variational inference
  - ...
- Key: exploit independencies of BN $\rightarrow$ d-separation property

Can X influence Y via Z?

Consider variables X, Y, and Z. Example model: flip coin, add result to sum of parents.

Network	Z latent	Z observed
Indirect causal effect 	Active $(X \not\perp Y)$	Not active $(X \perp Y \mid Z)$
Indirect evidential effect 	Active $(X \not\perp Y)$	Not active $(X \perp Y \mid Z)$
Common cause 	Active $(X \not\perp Y)$	Not active $(X \perp Y \mid Z)$
Common effect 	Not active $(X \perp Y)$	active $(X \not\perp Y \mid Z)$

d-separation

Definition

Let $\mathcal{G}$ be a BN structure and $X_1 \leftrightharpoons \ldots \leftrightharpoons X_n$ be a trail in $\mathcal{G}$. Denote by $\mathcal{E}$ the set of observed variables. The trail $X_1 \leftrightharpoons \ldots \leftrightharpoons X_n$ is **active** given $\mathcal{E}$ if

- Whenever we have a v-structure $X_{i-1} \to X_i \leftarrow X_{i+1}$ (common effect), then X_i or one of its descendants at in $\mathcal{Z}$, and
- no other node along the trail is in $\mathcal{Z}$.

Definition

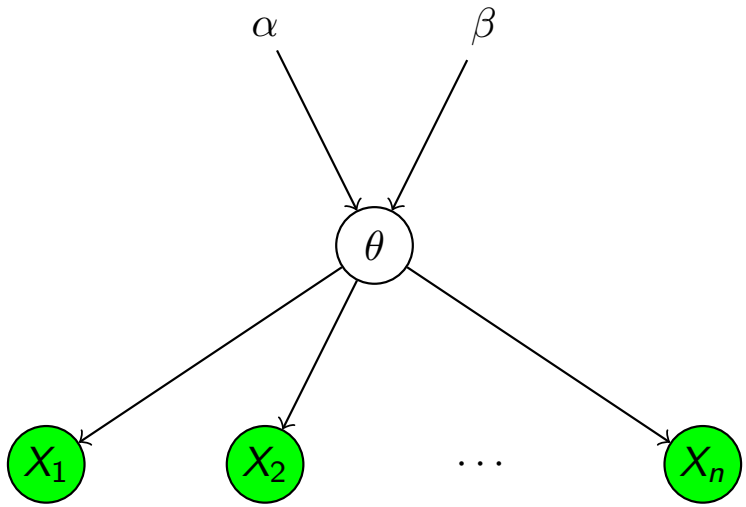
Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be three sets of vertices in $\mathcal{G}$. We say that $\mathcal{X}$ and $\mathcal{Y}$ are **d-separated** given $\mathcal{Z}$, denoted d-sep$(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})$, if there is no active trail between any node $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ given $\mathcal{Z}$.

Theorem (soundness)

If $\mathbb{P}$ *factorizes over* $\mathcal{G}$ *and* d-sep$(\mathcal{X}; \mathcal{Y} \mid \mathcal{Z})$, *then* $(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z})$.

Plate notation

- Suppose we observe the result $X_1, \ldots, X_n$ of n independent coin flips
- We want to infer the probability of heads θ
- Generative model
  - Prior: $\theta \sim \text{Beta}(\alpha, \beta)$
  - Flips: for all i, $X_i \sim \text{Bernoulli}(\theta)$
  - α, β are hyperparameters (fixed); θ is latent; $X_1, \ldots, X_n$ is observed
- **Plate notation** is a shortcut for "repeated" variables/subgraphs
  - Can be stacked (nested repeats)
  - Can be overlapping (all combinations of multiple indices)



Standard notation

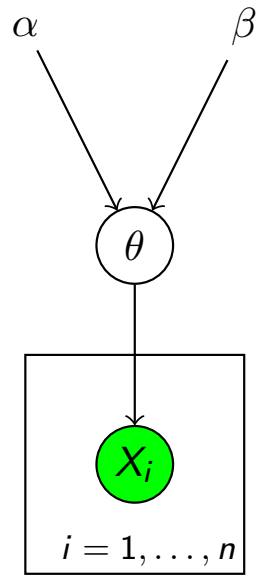


Plate notation

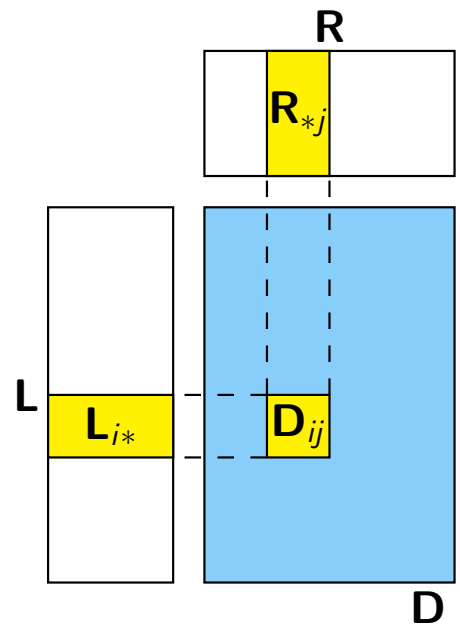
Outline

1. Background: Bayesian Networks
2. Probabilistic Matrix Factorization
3. Latent Dirichlet Allocation
4. Summary

Recap: Latent factor models

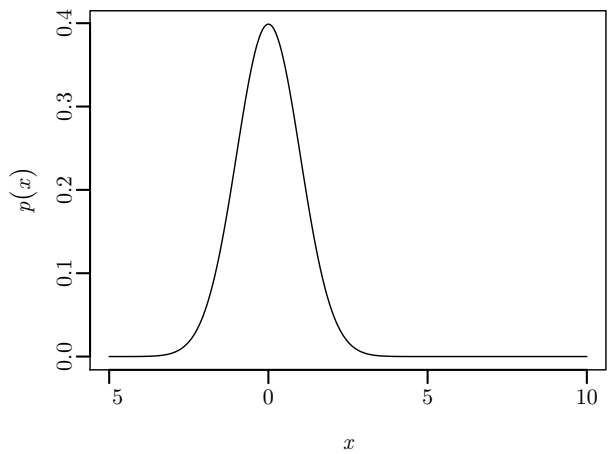
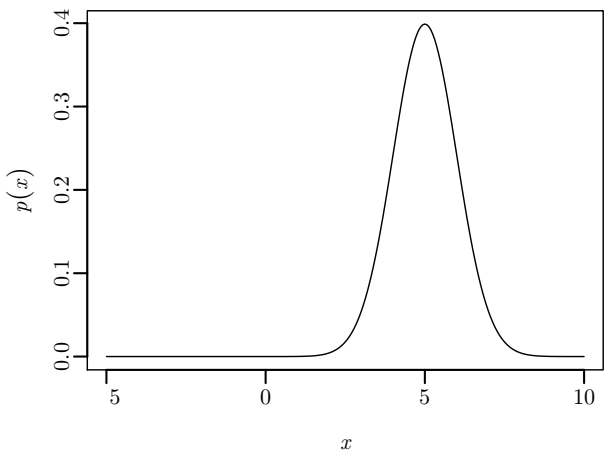
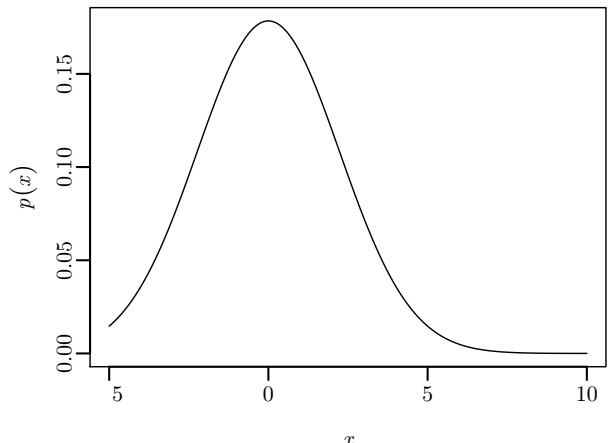
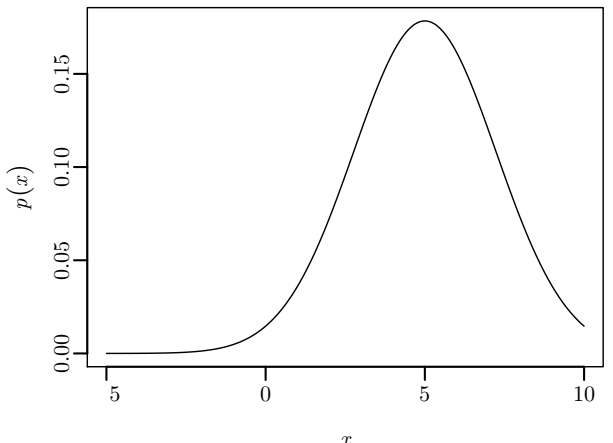
- m users, n items, $m \times n$ rating matrix $\mathbf{D}$
- Revealed entries $\Omega = \{ (i,j) \mid \text{rating } \mathbf{D}_{ij} \text{ is revealed} \}$
- User factors $\mathbf{L}_{m \times r}$, movie factors $\mathbf{R}_{r \times n}$
- Objective: $\text{argmin}_{\mathbf{L},\mathbf{R}} \sum_{(i,j) \in \Omega} \left[(\mathbf{D}_{ij} - [\mathbf{LR}]_{ij})^2 + \lambda_{\mathbf{L}} \|\mathbf{L}\|_F^2 + \lambda_{\mathbf{R}} \|\mathbf{R}\|_F^2 \right]$
- Prediction: $\hat{\mathbf{D}}_{ij} = \mathbf{L}_{i*}\mathbf{R}_{*j} = [\mathbf{LR}]_{ij}$

		R		
		Avatar (2.24)	*The Matrix* (1.92)	*Up* (1.18)
L	*Alice* (1.98)	? (4.4)	**4** (3.8)	**2** (2.3)
	Bob (1.21)	**3** (2.7)	**2** (2.3)	? (1.4)
	Charlie (2.30)	**5** (5.2)	? (4.4)	**3** (2.7)



Recap: Normal distribution (Gaussian distribution)

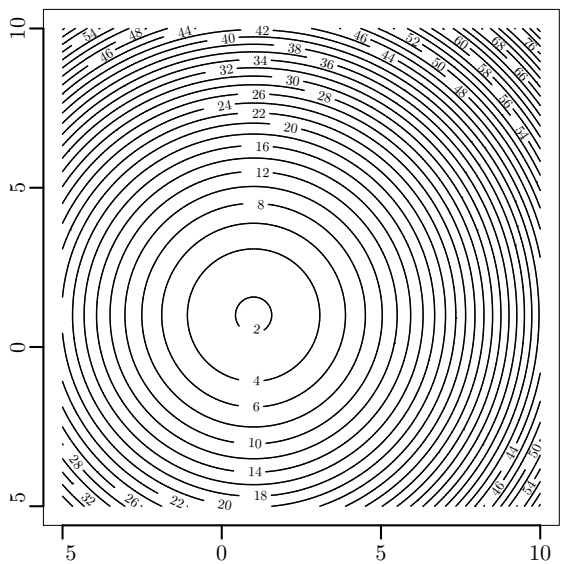
- Mean $\mu \in \mathbb{R}$, variance $\sigma^2 \in \mathbb{R}$ (or precision $\lambda = 1/\sigma^2$)
- Denoted Normal(μ, σ^2)
- Probability density function: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$

	$\mu = 0$	$\mu = 5$
$\sigma^2 = 1$		
$\sigma^2 = 5$		

Recap: Multivariate normal distribution

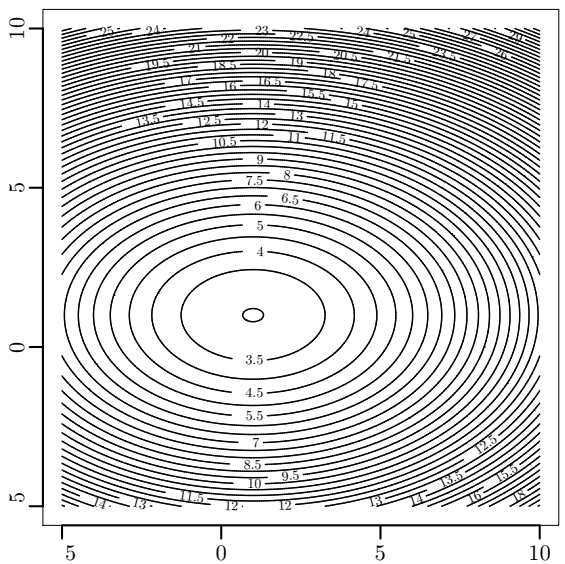
- Mean $\boldsymbol{\mu} \in \mathbb{R}^k$, covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ (or precision $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$)
- Denoted Normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Let $|\boldsymbol{\Sigma}|$ be the determinant of $\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ is positive definite:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$



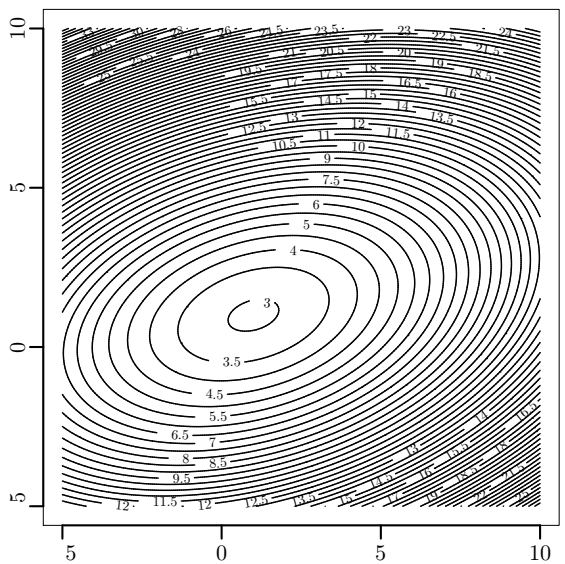
$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \sigma^2 \mathbf{I}$$

(spherical)



$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$$

(diagonal)

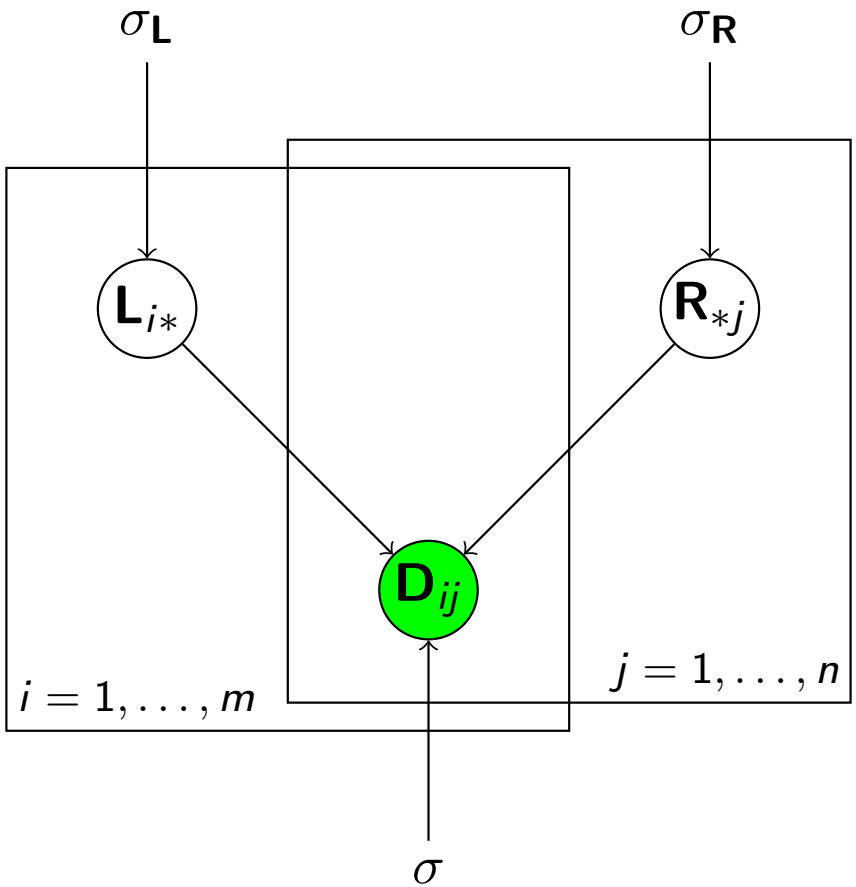


$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}$$

Probabilistic linear model with Gaussian noise (PMF)

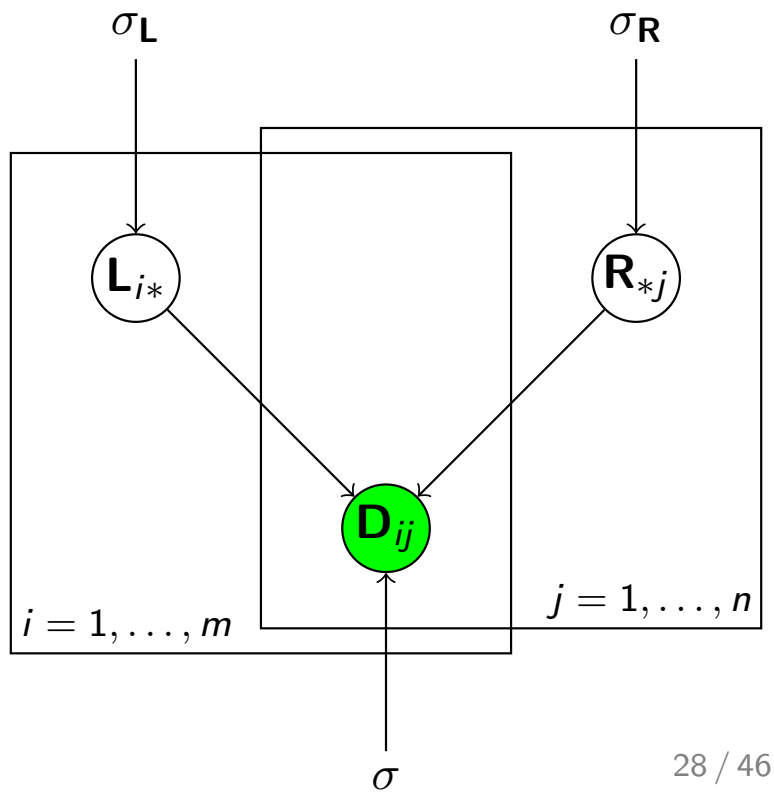
Hyperparameters: $\sigma_{\mathbf{L}}$ (sd of entries of $\mathbf{L}$), $\sigma_{\mathbf{R}}$ (sd of $\mathbf{R}$), σ (sd of noise)

1. For each user i, draw $\mathbf{L}_{i*}$ from Normal$(0, \sigma_{\mathbf{L}}^2 \mathbf{I})$
2. For each movie j, draw $\mathbf{R}_{*j}$ from Normal$(0, \sigma_{\mathbf{R}}^2 \mathbf{I})$
3. For each rating (i, j), draw $\mathbf{D}_{ij}$ from Normal$([\mathbf{LR}]_{ij}, \sigma^2)$



Let's analyze: posterior distribution

$$
\begin{aligned}
& p(\mathbf{L}, \mathbf{R} \mid \mathbf{D}, \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2) \\
&= \frac{p(\mathbf{L}, \mathbf{R}, \mathbf{D} \mid \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2)}{p(\mathbf{D} \mid \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2)} \\
&\propto p(\mathbf{L}, \mathbf{R}, \mathbf{D} \mid \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2) \\
&= p(\mathbf{D} \mid \mathbf{L}, \mathbf{R}, \sigma^2) p(\mathbf{L} \mid \sigma_{\mathbf{L}^2}) p(\mathbf{R} \mid \sigma_{\mathbf{R}}^2) \\
&\propto \left[\prod_{(i,j) \in \Omega} \exp\left(-\frac{(\mathbf{D}_{ij} - [\mathbf{LR}]_{ij})^2}{2\sigma^2} \right) \right] \\
&\quad \cdot \left[\prod_{i,k} \exp\left(-\frac{\mathbf{L}_{ik}^2}{2\sigma_{\mathbf{L}}^2} \right) \right] \\
&\quad \cdot \left[\prod_{k,j} \exp\left(-\frac{\mathbf{R}_{ik}^2}{2\sigma_{\mathbf{R}}^2} \right) \right]
\end{aligned}
$$



Let's analyze: MAP estimate

$$
\begin{aligned}
&\text{MAP}(\mathbf{L}, \mathbf{R} \mid \mathbf{D}, \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2) \\
&= \operatorname*{argmax}_{\mathbf{L},\mathbf{R}} p(\mathbf{L}, \mathbf{R}, \mid \mathbf{D}, \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2) \\
&= \operatorname*{argmin}_{\mathbf{L},\mathbf{R}} -\ln p(\mathbf{L}, \mathbf{R} \mid \mathbf{D}, \sigma^2, \sigma_{\mathbf{L}}^2, \sigma_{\mathbf{R}}^2) \\
&= \operatorname*{argmin}_{\mathbf{L},\mathbf{R}} \frac{1}{2\sigma^2} \sum_{(i,j)\in\Omega} (\mathbf{D}_{ij} - [\mathbf{LR}]_{ij})^2 + \frac{1}{2\sigma_{\mathbf{L}}^2} \sum_{i,k} \mathbf{L}_{ik}^2 + \frac{1}{2\sigma_{\mathbf{R}}^2} \sum_{k,j} \mathbf{R}_{kj}^2 \\
&= \operatorname*{argmin}_{\mathbf{L},\mathbf{R}} \sum_{(i,j)\in\Omega} (\mathbf{D}_{ij} - [\mathbf{LR}]_{ij})^2 + \frac{\sigma^2}{\sigma_{\mathbf{L}}^2} \sum_{i,k} \mathbf{L}_{ik}^2 + \frac{\sigma^2}{\sigma_{\mathbf{R}}^2} \sum_{k,j} \mathbf{R}_{kj}^2 \\
&= \operatorname*{argmin}_{\mathbf{L},\mathbf{R}} \sum_{(i,j)\in\Omega} (\mathbf{D}_{ij} - [\mathbf{LR}]_{ij})^2 + \lambda_{\mathbf{L}} \|\mathbf{L}\|_F^2 + \lambda_{\mathbf{R}} \|\mathbf{R}\|_F^2
\end{aligned}
$$

- PMF + MAP = latent factor model with L2 regularization
- Precision $\lambda_{\mathbf{L}} = \sigma^2/\sigma_{\mathbf{L}}^2$ relates variation of noise and factors
- Similarly $\lambda_{\mathbf{R}} = \sigma^2/\sigma_{\mathbf{R}}^2$

Did we achieve anything?

- MAP estimate does not allow us to judge uncertainty individually for each prediction
  - Pick $(i, j) \notin \Omega$
  - By assumption, noise is i.i.d. Normal$(0, \sigma^2)$ *given* $\mathbf{L}$ and $\mathbf{R}$

    $$\hat{\mathbf{D}}_{ij} \sim \text{Normal}([\mathbf{LR}]_{ij}, \sigma^2)$$

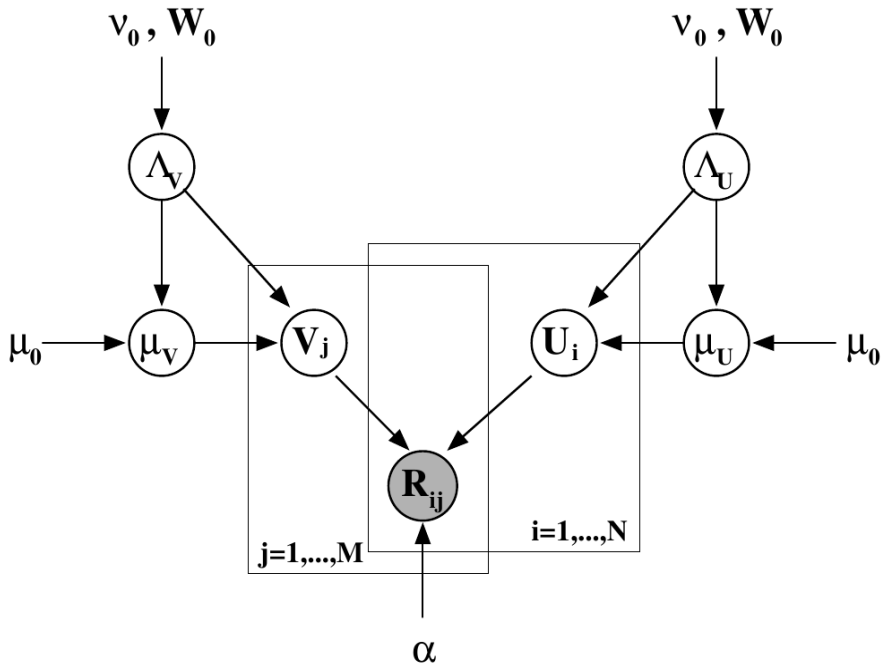
  - With PMF, we can marginalize out $\mathbf{L}$ and $\mathbf{R}$.

    $$\begin{aligned} & p(\hat{\mathbf{D}}_{ij} \mid \mathbf{D}, \sigma^2, \sigma^2_{\mathbf{L}}, \sigma^2_{\mathbf{R}}) \\ &= \int_{\mathbf{L},\mathbf{R}} p(\hat{\mathbf{D}}_{ij} \mid \mathbf{L}, \mathbf{R}, \sigma^2) p(\mathbf{L}, \mathbf{R} \mid \mathbf{D}, \sigma^2, \sigma^2_{\mathbf{L}}, \sigma^2_{\mathbf{R}}) \, \mathrm{d}\mathbf{L} \, \mathrm{d}\mathbf{R} \\ &= \int_{\mathbf{L},\mathbf{R}} p_{\text{Normal}}(\hat{\mathbf{D}}_{ij} \mid [\mathbf{LR}]_{ij}, \sigma^2) p(\mathbf{L}, \mathbf{R} \mid \mathbf{D}, \sigma^2, \sigma^2_{\mathbf{L}}, \sigma^2_{\mathbf{R}}) \, \mathrm{d}\mathbf{L} \, \mathrm{d}\mathbf{R} \end{aligned}$$

  - We obtain a "customized" distribution for $\hat{\mathbf{D}}_{ij}$

- Better understanding of latent factor models
  - Probabilistic models reveal underlying assumptions
  - Easier to play with assumptions or integrate additional data points

Example: Bayesian prob. matrix factorization (BPFM)



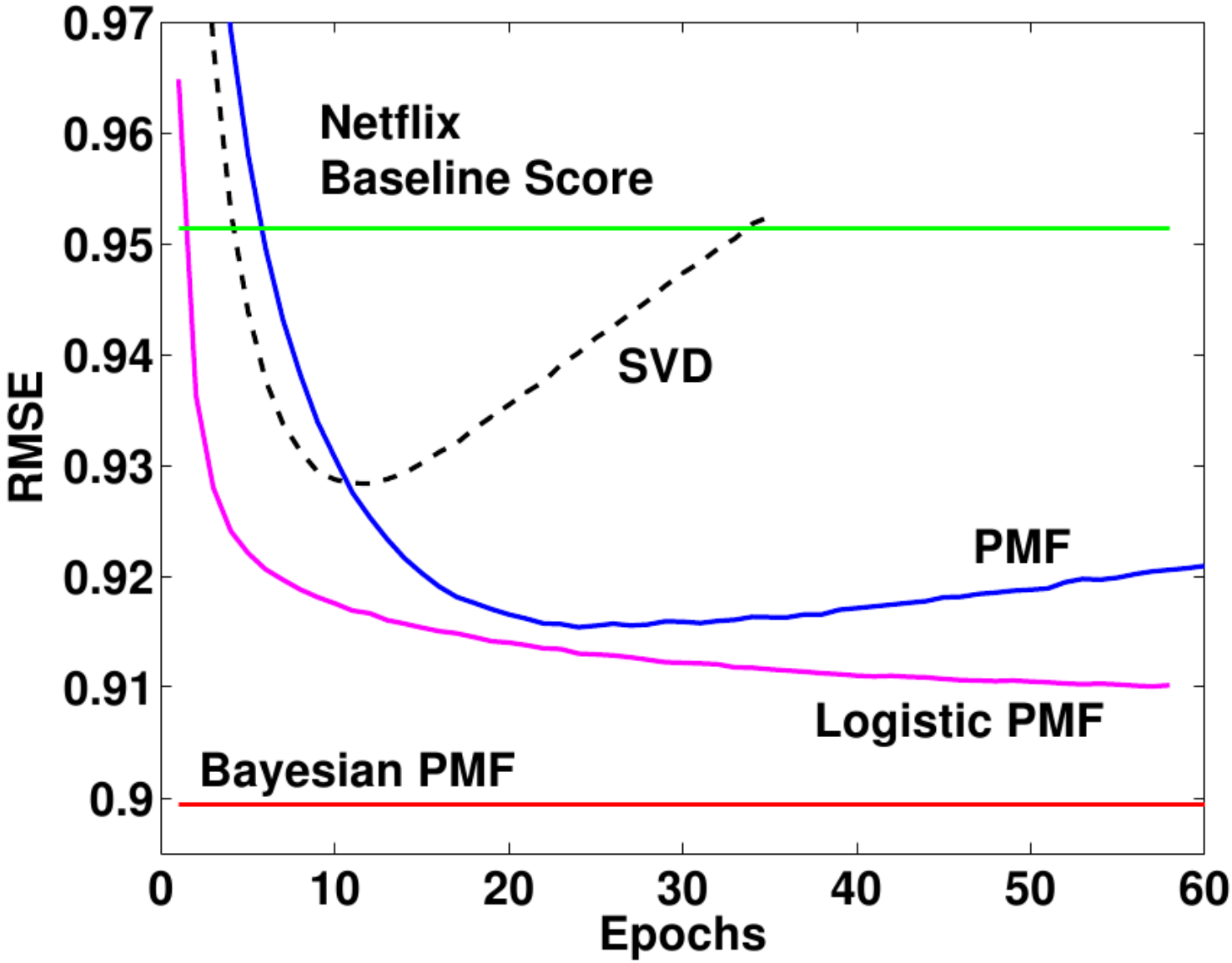
$$(\mathbf{L} = \mathbf{U}^T, \mathbf{R} = \mathbf{V})$$

Goal: automatic complexity control

$\rightarrow$ Model mean, variance, and covariance of factors

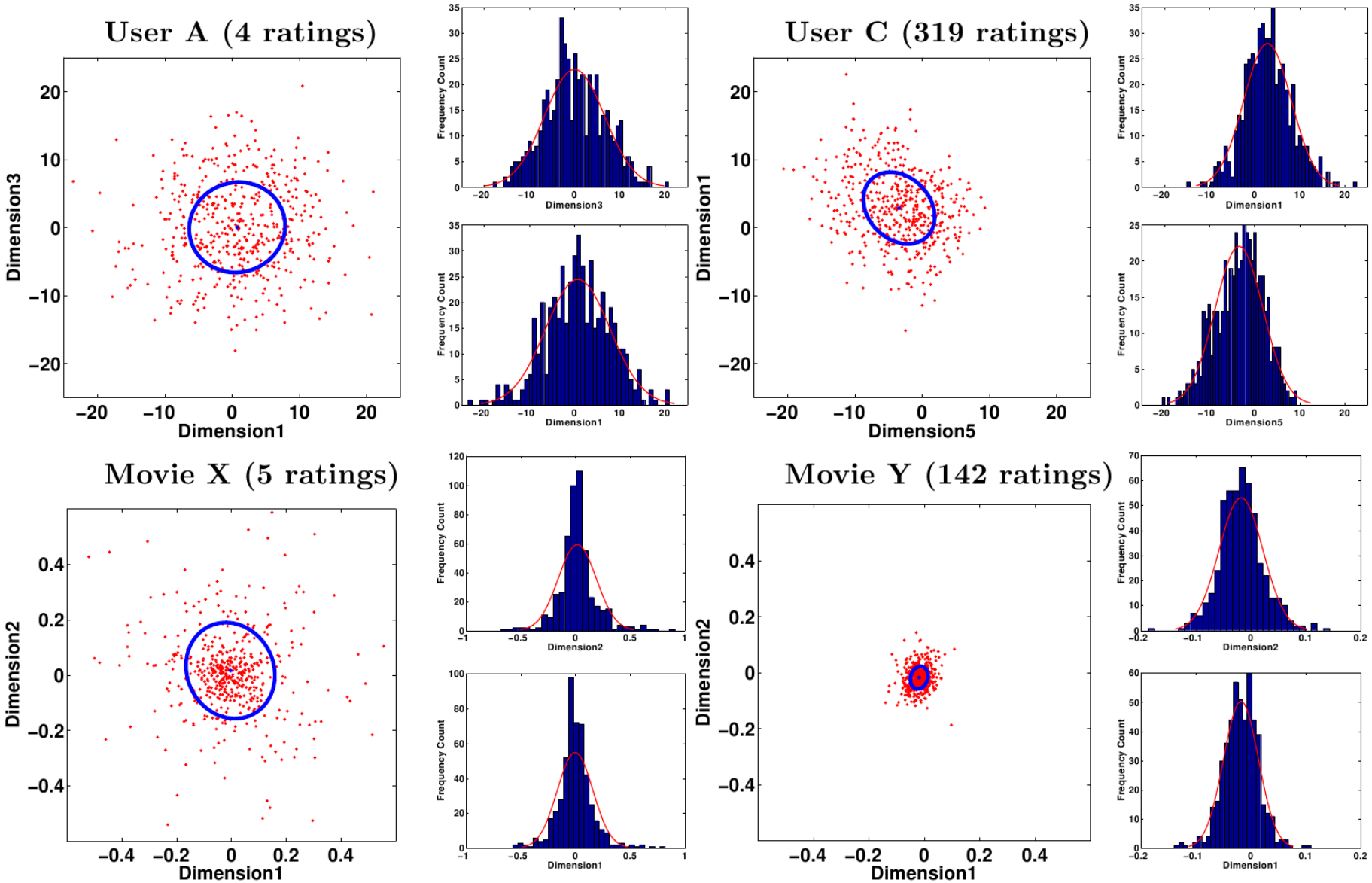
1. Sample precision matrix for users ($\Lambda_\mathbf{U}$) and movies ($\Lambda_\mathbf{V}$)
2. Sample factor means: e.g., $\mu_\mathbf{V} \sim \text{Normal}(\mu_0, (\beta_0 \Lambda_\mathbf{V})^{-1})$
3. Sample factors: $\mathbf{V}_j \sim \text{Normal}(\mu_\mathbf{V}, \Lambda_\mathbf{V}^{-1})$
4. Sample ratings from $\text{Normal}([\mathbf{U}^T\mathbf{V}]_{ij}, \alpha^{-1})$

BPMF: quality on validation data

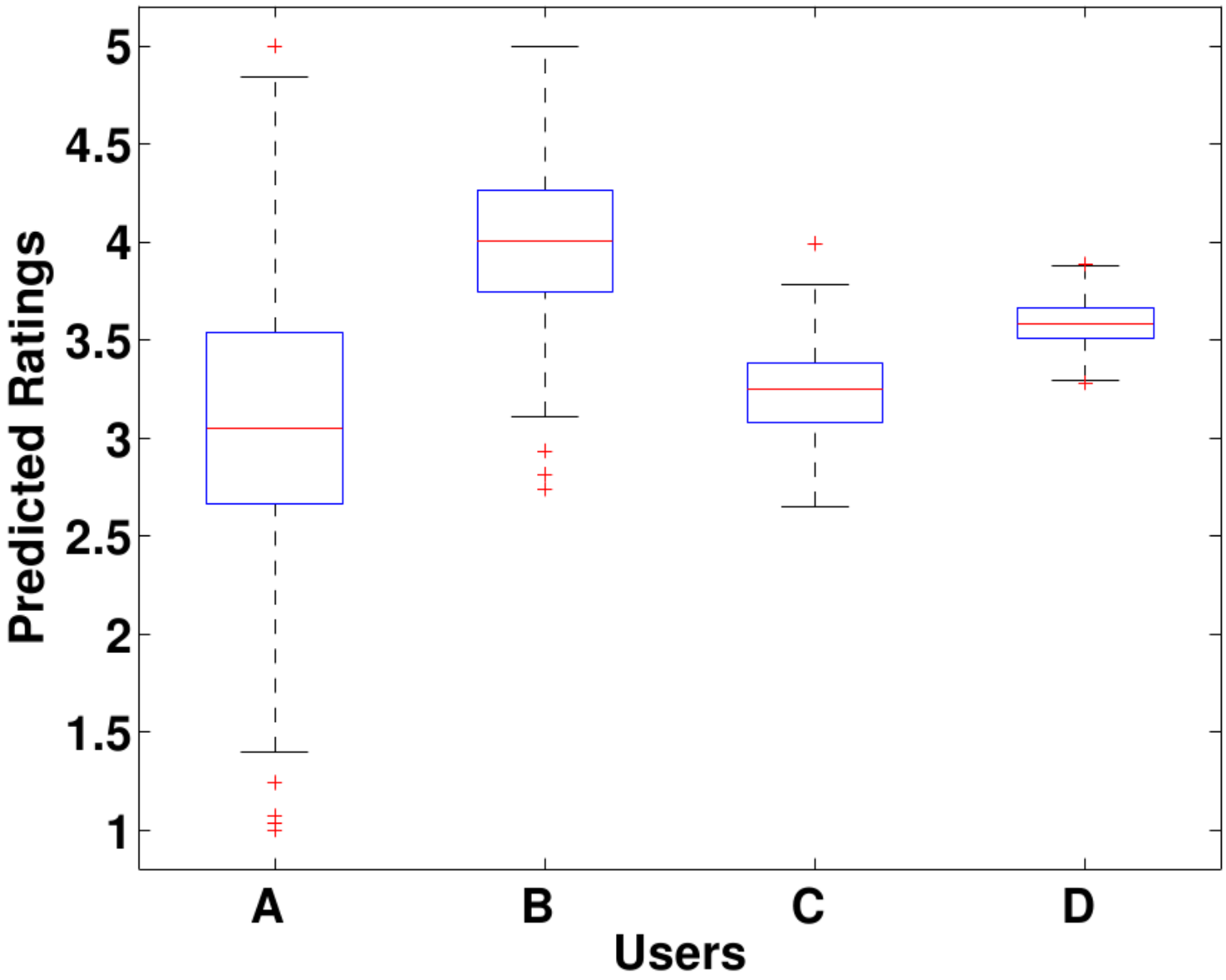


Salakhutdinov and Mnih, 2008

BPMF: Example (1)



BPMF: Example (2)



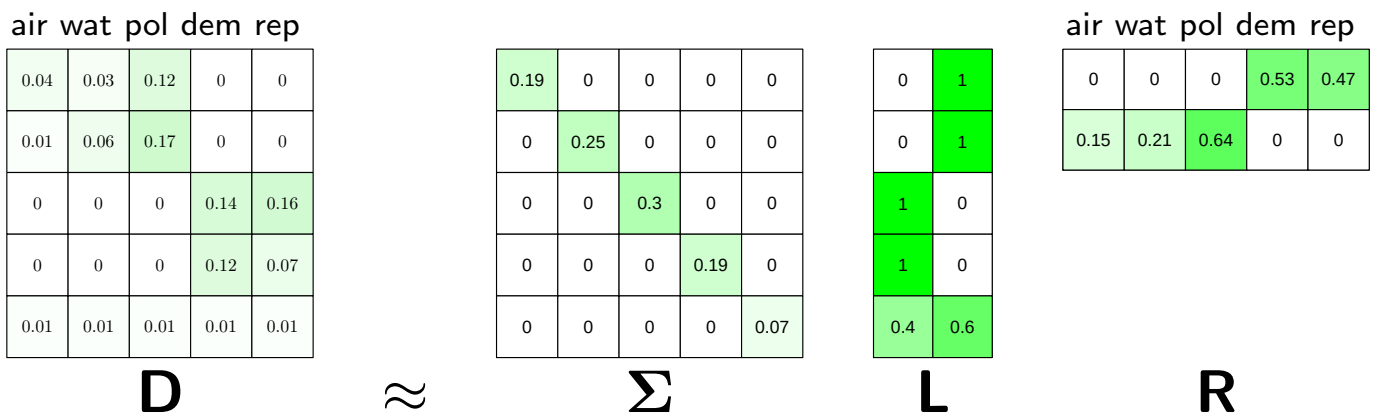
(A, B, C, D have 4, 24, 319, 660 ratings, respectively)

Outline

1. Background: Bayesian Networks
2. Probabilistic Matrix Factorization
3. Latent Dirichlet Allocation
4. Summary

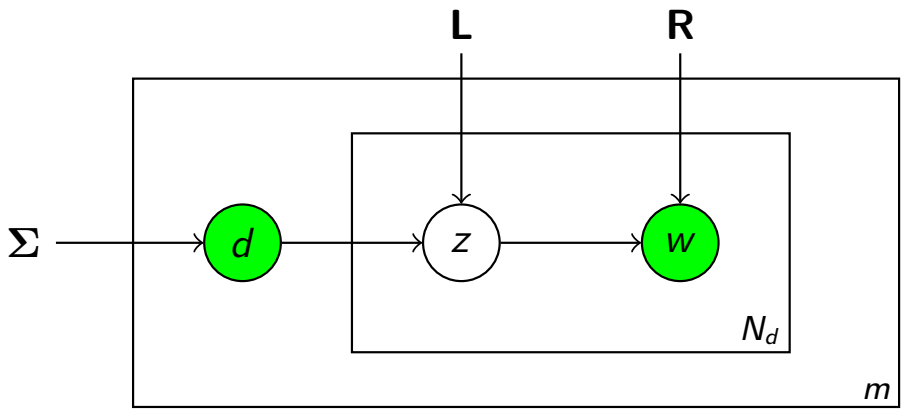
Recap: Probabilistic latent semantic analysis (pLSA)

- **D** is an $m \times n$ document-word matrix (normalized to sum to 1)
- pLSA reveals topics by factoring $\mathbf{D} \approx \mathbf{\Sigma L R}$, where
  - $\mathbf{\Sigma}$ is an $m \times m$ diagonal matrix
    $\rightarrow$ Document probabilities
  - **L** is an $m \times r$ row-stochastic matrix (rows sum to 1)
    $\rightarrow$ Topic mixture per document
  - **R** is an $r \times n$ row-stochastic matrix (rows sum to 1)
    $\rightarrow$ Word distribution per topic



pLSA as a generative model

- Generating a word
  1. Select document $d = d_i$ with probability $\mathbb{P}(d_i) = \mathbf{\Sigma}_{ii}$:
  2. Select topic $z = z_k$ with probability $\mathbb{P}(z_k \mid d) = \mathbf{L}_{dk}$
  3. Generate word $w = w_j$ with probability $\mathbb{P}(w_j \mid z) = \mathbf{R}_{zj}$
- Alternative way to write this
  1. $d \sim \text{Multinomial}(\text{diag}(\mathbf{\Sigma}), 1)$
  2. $z \sim \text{Multinomial}(\mathbf{L}_{d*}, 1)$
  3. $w \sim \text{Multinomial}(\mathbf{R}_{z*}, 1)$



	air	wat	pol	dem	rep
	0.04	0.03	0.12	0	0
	0.01	0.06	0.17	0	0
	0	0	0	0.14	0.16
	0	0	0	0.12	0.07
	0.01	0.01	0.01	0.01	0.01

D

$\approx$

0.19	0	0	0	0
0	0.25	0	0	0
0	0	0.3	0	0
0	0	0	0.19	0
0	0	0	0	0.07

Σ

0	1
0	1
1	0
1	0
0.4	0.6

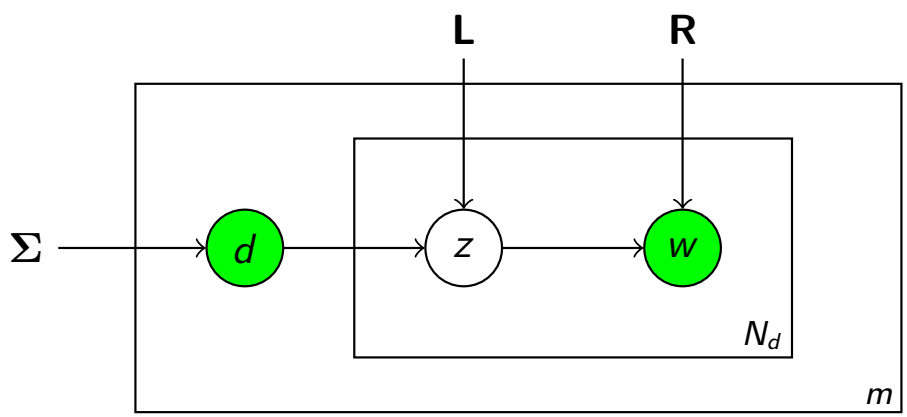
L

air	wat	pol	dem	rep
0	0	0	0.53	0.47
0.15	0.21	0.64	0	0

R

Problems with pLSA

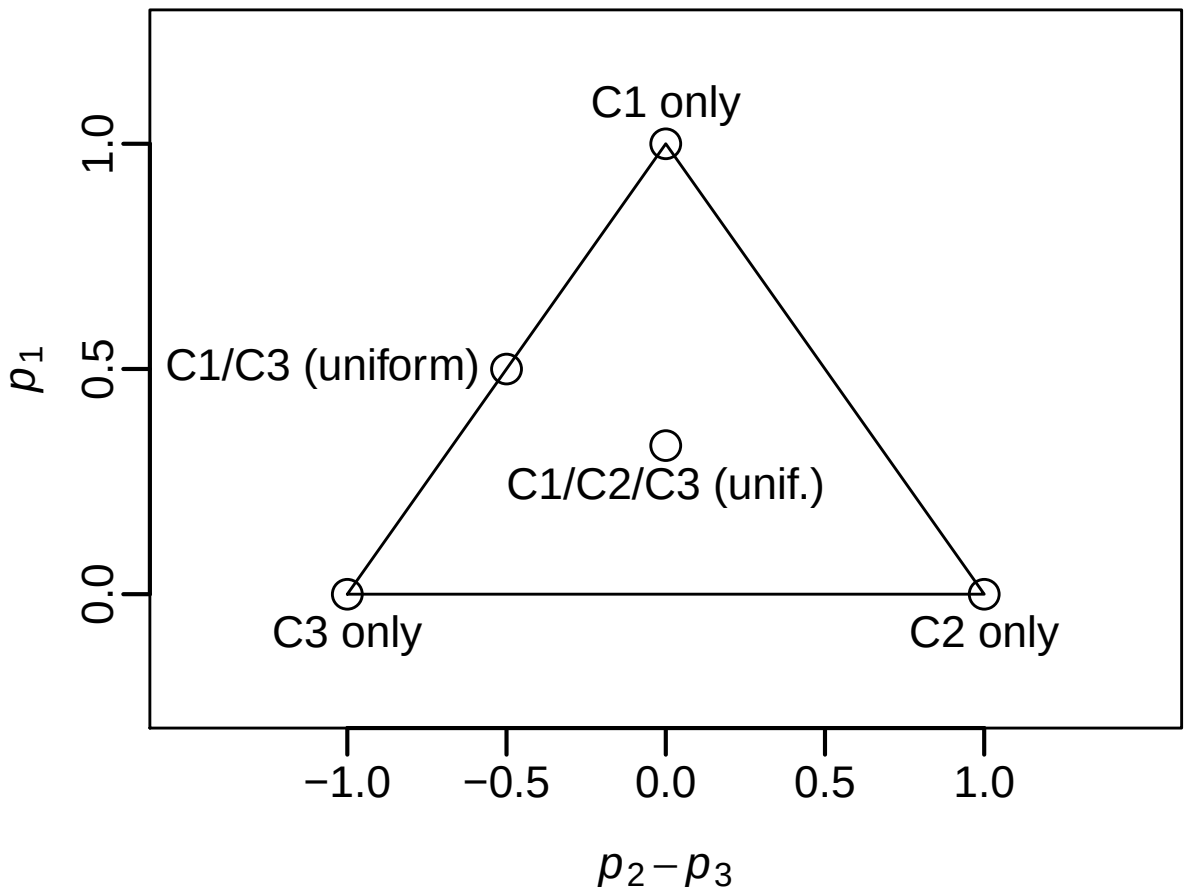
- Not a well-defined generative model of documents
  - Learns m mixtures (=rows of $\mathbf{L}$) $\rightarrow$ m possible values for d
  - Not clear how handle documents outside of the training set
    $\rightarrow$ A "fold-in" heuristic is often used
  - Number of parameters grows linearly with number of documents
    ($mr + nr$ mixture parameters)
    $\rightarrow$ Leads to overfitting (reduced via "tempering")
- "No" priors on document-topic ($\mathbf{L}$) or topic-word distributions ($\mathbf{R}$)
  - One can show: pLSA related to MAP estimate of LDA with uniform Dirichlet prior
- Latent Dirichlet allocation (LDA) addresses these problems



Dirichlet distribution (1)

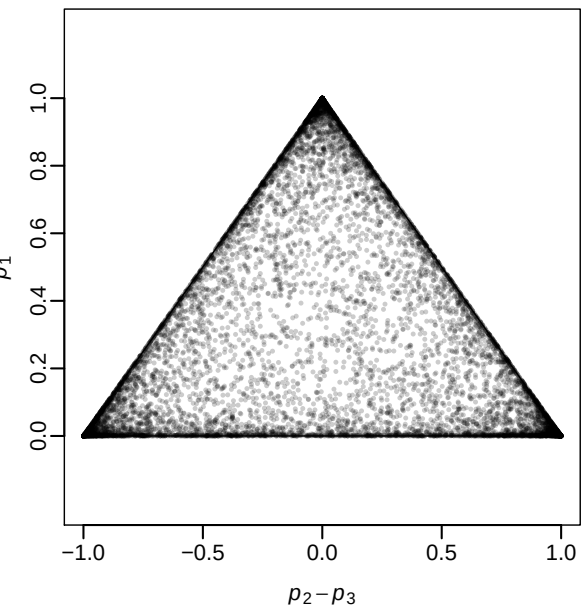
- Conjugate prior for the multinomial distribution over K categories
- Distribution over vectors $\mathbf{p} \in \mathbb{R}_+^K$ satisfying $\|\mathbf{p}\|_1 = \sum_k p_k = 1$
- $\mathbf{p}$ can be seen as parameters of a multinomial distribution:

$$\mathbf{p}_k = \text{probability to select category } k \text{ (in one trial)}$$

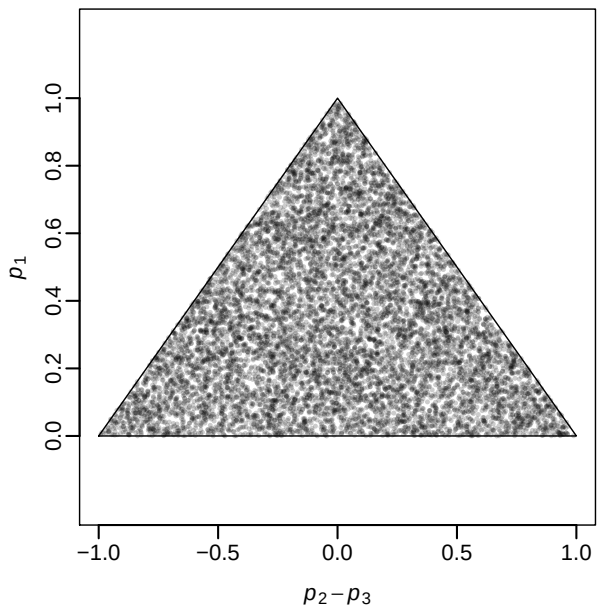


Dirichlet distribution (2)

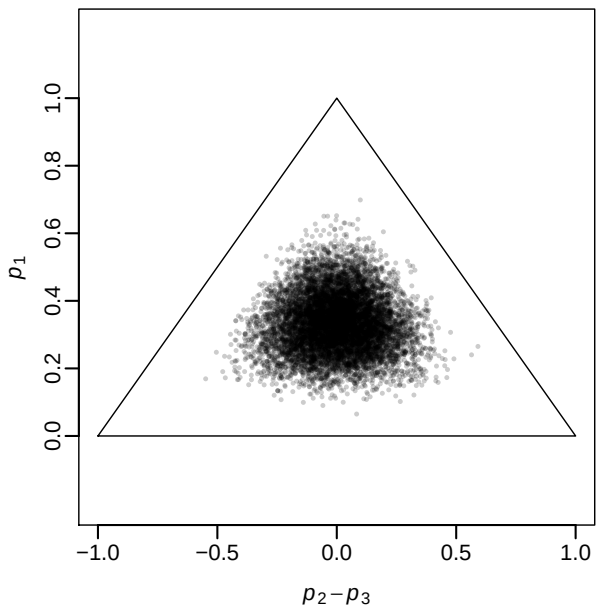
- Parameterized by vector $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ with $\boldsymbol{\alpha}_k > 0$ ("concentration parameters")
- $p(\mathbf{x} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \mathbf{x}_k^{\boldsymbol{\alpha}_k - 1}$
- Special case: **symmetric Dirichlet distribution**
  - Single **concentration parameter** α; set $\boldsymbol{\alpha}_k = \alpha$
  - $\alpha \ll 1$: multinomials concentrate around single category (sparse)
  - $\alpha \gg 1$: multinomials spread uniformly over categories (dense)
  - $\alpha = 1$: uniform distribution over multinomials



$\alpha = .3$



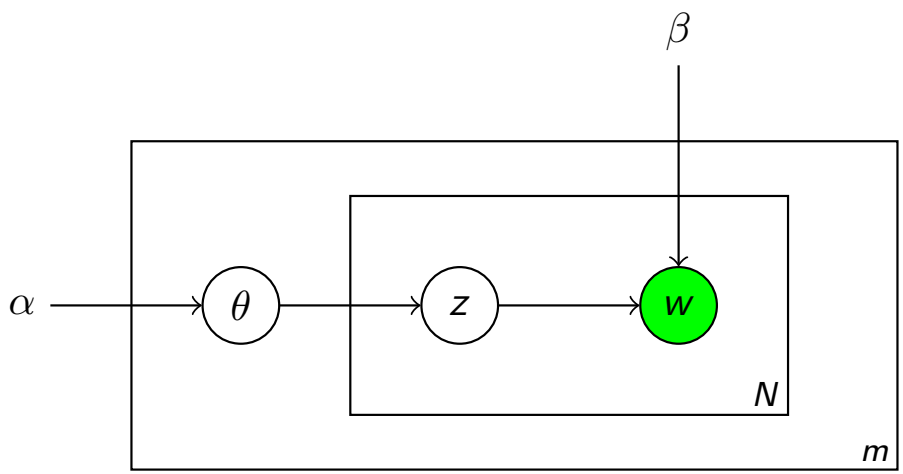
$\alpha = 1$



$\alpha = 10$

Latent Dirichlet Allocation (LDA)

- Parameters
  - $\xi \in \mathbb{R}_+$: mean number of words per document
  - $\alpha \in \mathbb{R}_+^r$: concentration parameter for topic mixture (usually $\alpha \ll 1$)
  - $\beta \in \mathbb{R}_+^{r \times n}$: word distribution for each topic
- For each document:
  1. Choose number of words $N \sim \text{Poisson}(\xi)$
  2. Choose topic mixture $\theta \sim \text{Dirichlet}(\alpha)$
  3. For each of the N words:
     1. Choose a topic $z_n \sim \text{Multinomial}(\theta, 1)$
     2. Choose a word $w_n \sim \text{Multinomial}(\beta_{z_n *}, 1)$

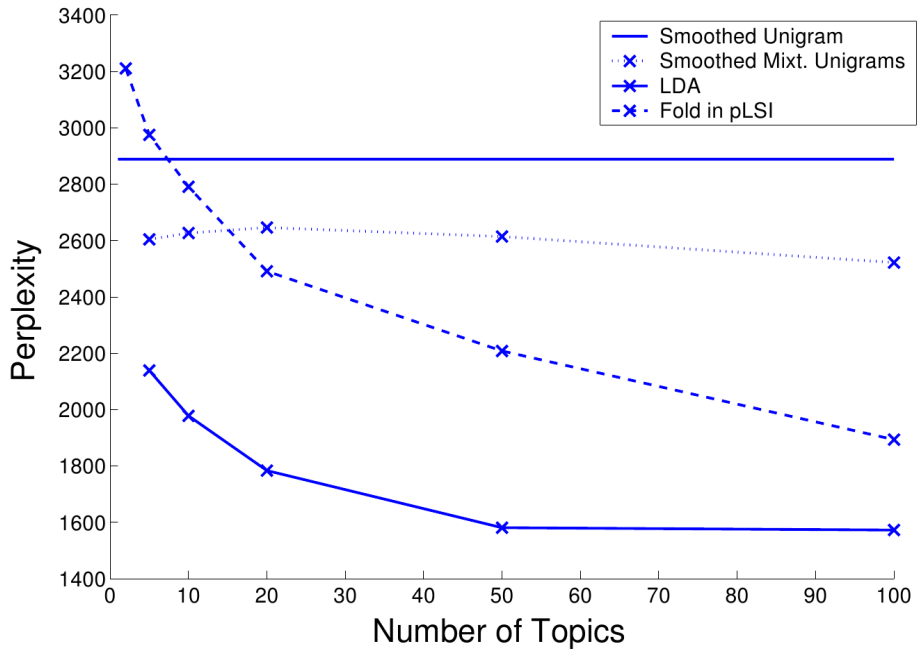


Is this better than pLSA?

- One way to measure: generalization performance on new documents
- **Perplexity** is a often used to measure generalization performance
- Test set $\mathbf{D}_{\text{test}}$ of m_{test} previously unseen documents

$$\text{perplexity}(\mathbf{D}_{\text{test}}) = \exp\left(-\frac{\sum_{d=1}^{m_{\text{test}}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{m_{\text{test}}} N_d}\right)$$

- Higher likelihood of test data $\rightarrow$ lower perplexity



5225 scientific abstracts (90% train, 10% test)

Blei et al., 2003.

LDA example

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

β

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

z

Blei et al., 2003.

Outline

1. Background: Bayesian Networks
2. Probabilistic Matrix Factorization
3. Latent Dirichlet Allocation
4. Summary

Lessons learned

- Bayesian networks
  - Models direct probabilistic interaction via a directed acyclic graph
  - Priors + model + data $\rightarrow$ posterior (via probabilistic inference)
  - Posterior captures belief about the values of latent variables
- Probabilistic matrix factorization (for collaborative filtering)
  - PMF + MAP inference = latent factor models with L2 regularization
  - Can be customized in various ways
  - Allows quantifying the uncertainty of each prediction
- Latent dirichlet allocation (for topic modelling)
  - Widely used generative model for text corpora
  - Addresses some limitations of pLSI
  - Many extensions exist (e.g., to add supervision or n-gram modelling)

Suggested reading

- Daphne Koller, Nir Friedman
  Probabilistic Graphical Models: Principles and Techniques (Ch. 3)
  The MIT Press, 2009

- Ruslan Salakhutdinov, Andriy Mnih
  Probabilistic Matrix Factorization
  Advances in Neural Information Processing Systems (NIPS), 2008
  `http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2007_1007.pdf`

- David M. Blei, Andrew Y. Ng, Michael I. Jordan
  Latent Dirichlet Allocation
  Journal of Machine Learning Research 3, 2003
  `http://dl.acm.org/citation.cfm?id=944937`