Intro. to Topic Modeling (cont’d) + Factor Analysis

Kayhan Batmanghelich

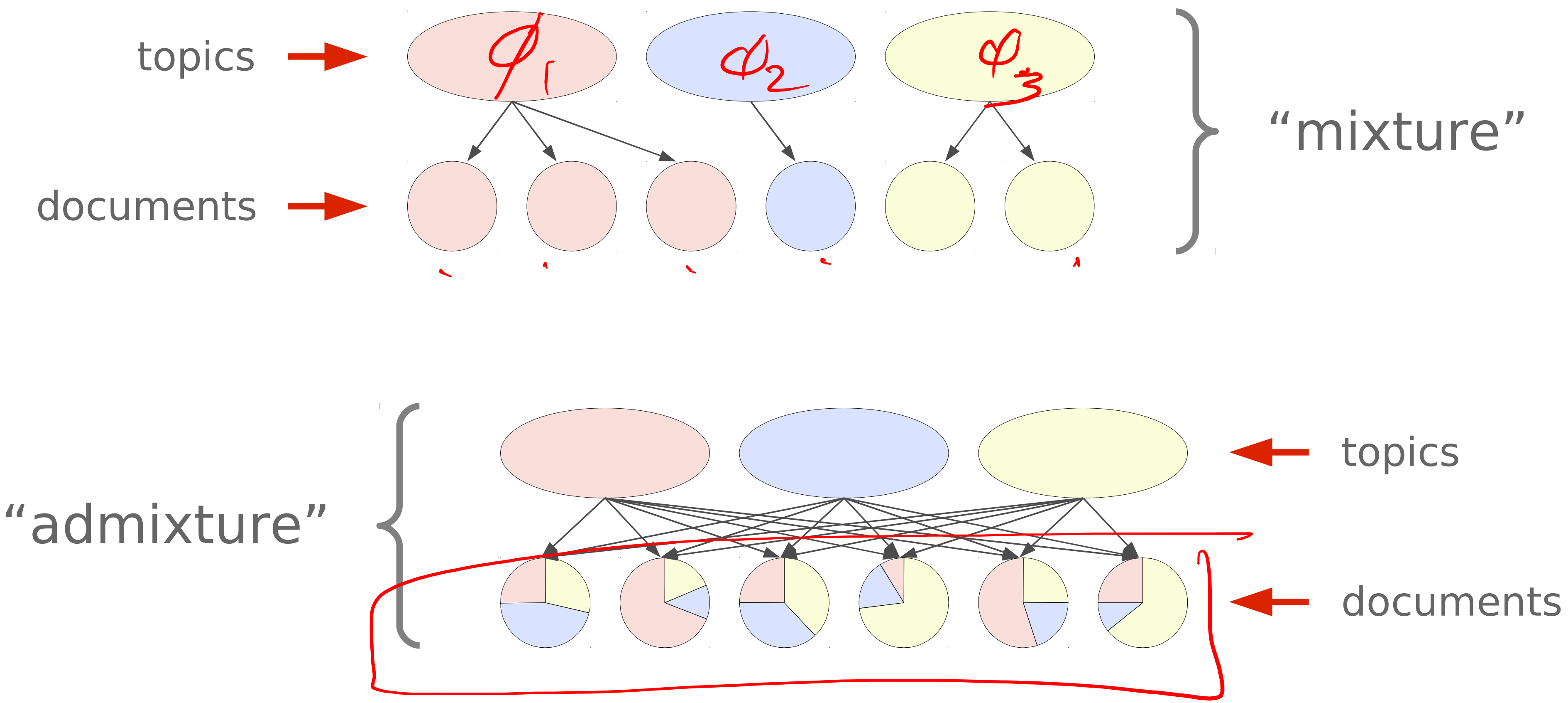
Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

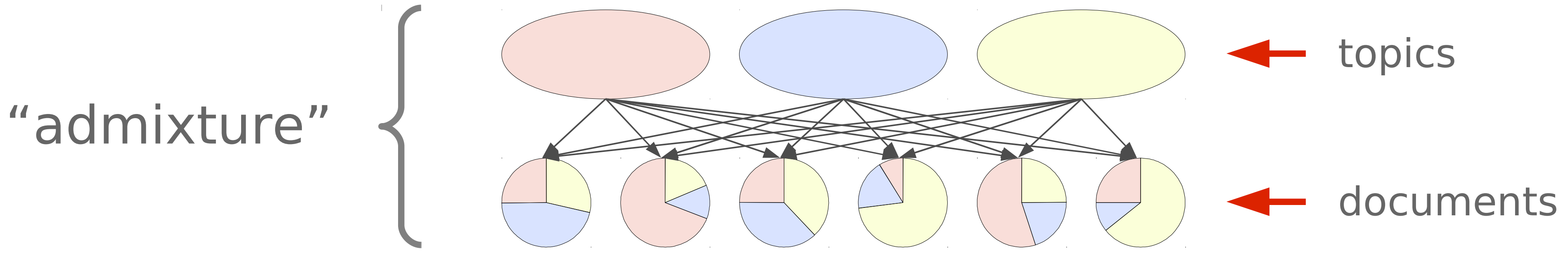
Mixture vs. Admixture (LDA)



Diagrams from Wallach, JHU 2011, slides

Latent Dirichlet Allocation

- Generative Process



- Example corpus

the	he	is	the	and	the	she	she	is	is
x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	x_{31}	x_{32}	x_{33}	x_{34}

Document 1 | Document 2 | Document 3

Latent Dirichlet Allocation

- Generative Process

For each topic $k \in \{1, \dots, K\}$:

$\phi_k \sim \text{Dir}(\boldsymbol{\beta})$ Dir *[draw distribution over words]*

For each document $m \in \{1, \dots, M\}$

$\boldsymbol{\theta}_m \sim \text{Dir}(\boldsymbol{\alpha})$ θ_k # topic *[draw distribution over topics]*

Docum Specif Dist of topic

For each word $n \in \{1, \dots, N_m\}$

$z_{mn} \sim \text{Mult}(1, \boldsymbol{\theta}_m)$ *[draw topic assignment]*

$x_{mn} \sim \phi_{z_{mi}}$ *[draw word]*

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

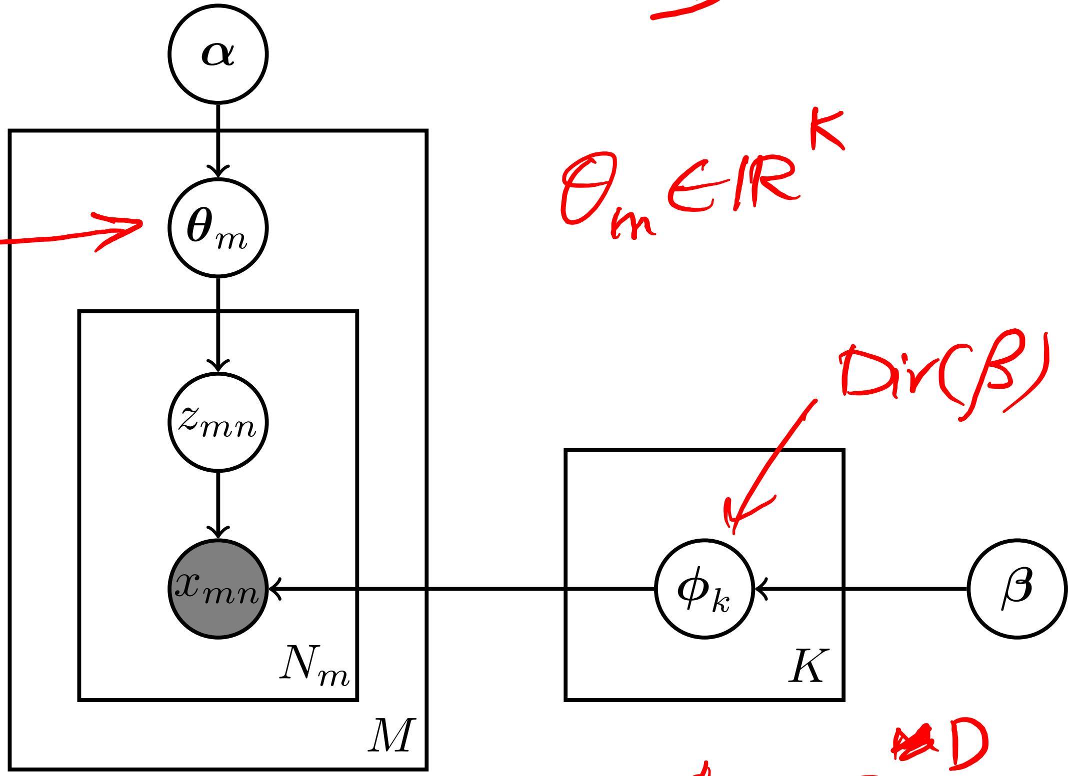
Latent Dirichlet Allocation

- Plate Diagram

K #topic

D dic (word)

Dir(α)



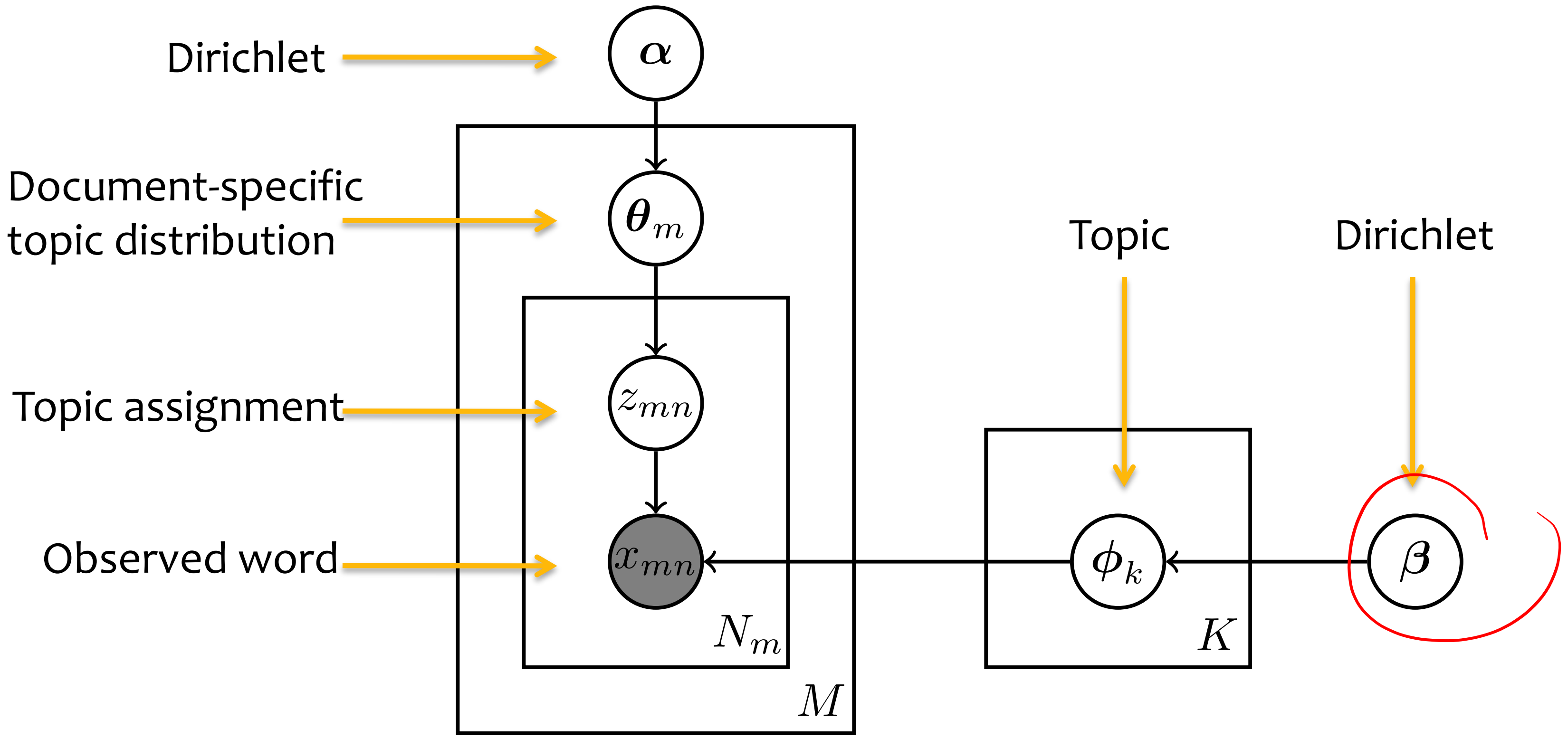
$\theta_m \in \mathbb{R}^K$

Dir(β)

$\phi_k \in \mathbb{R}^D$

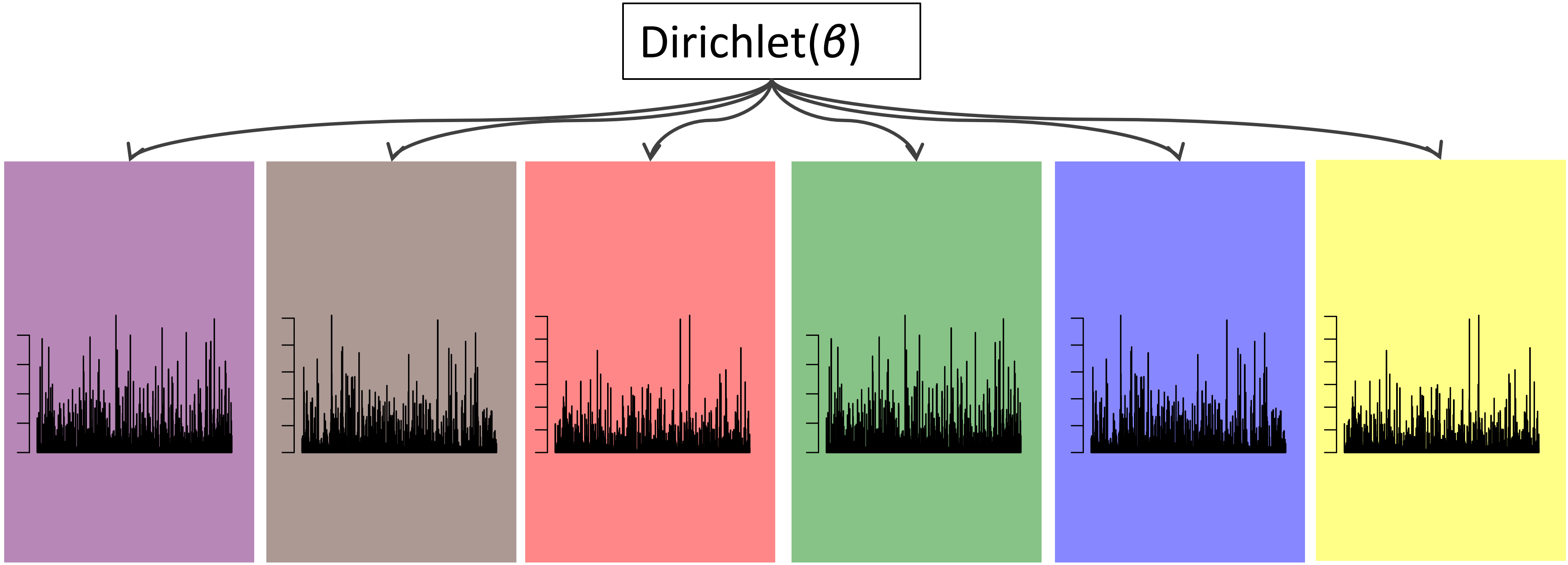
Latent Dirichlet Allocation

- Plate Diagram



(Blei, Ng, & Jordan, 2003)

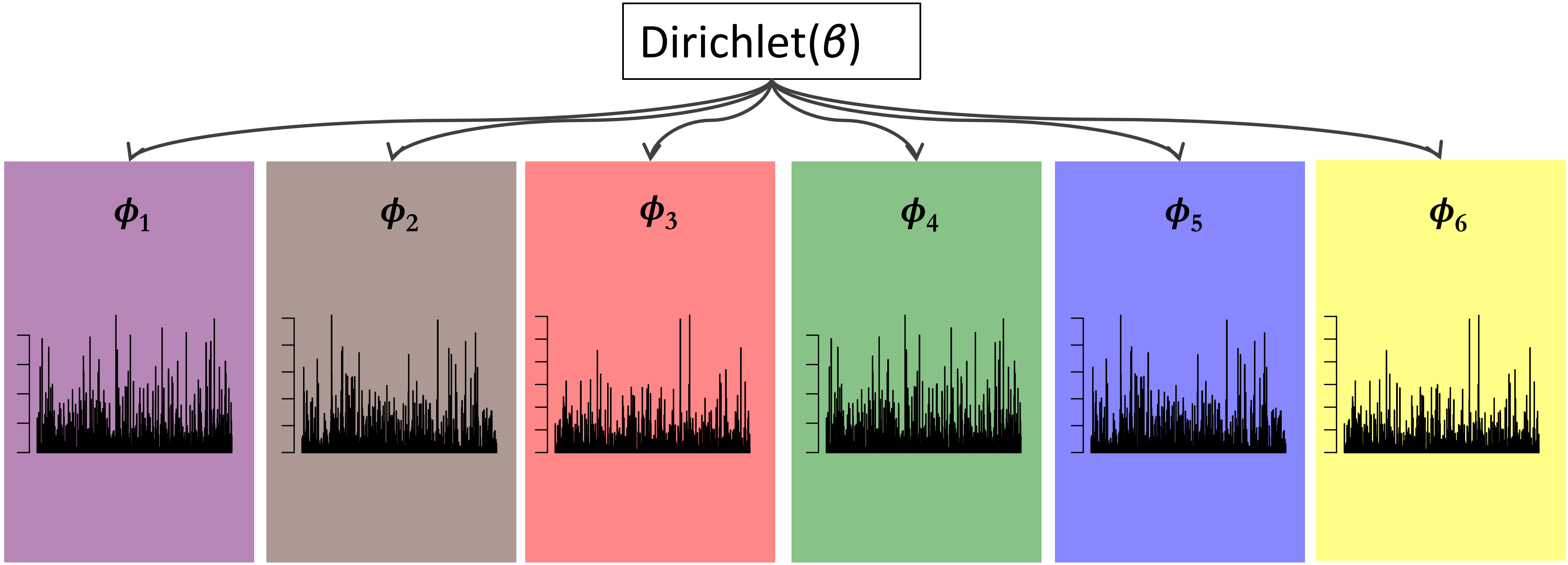
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_k$

(Blei, Ng, & Jordan, 2003)

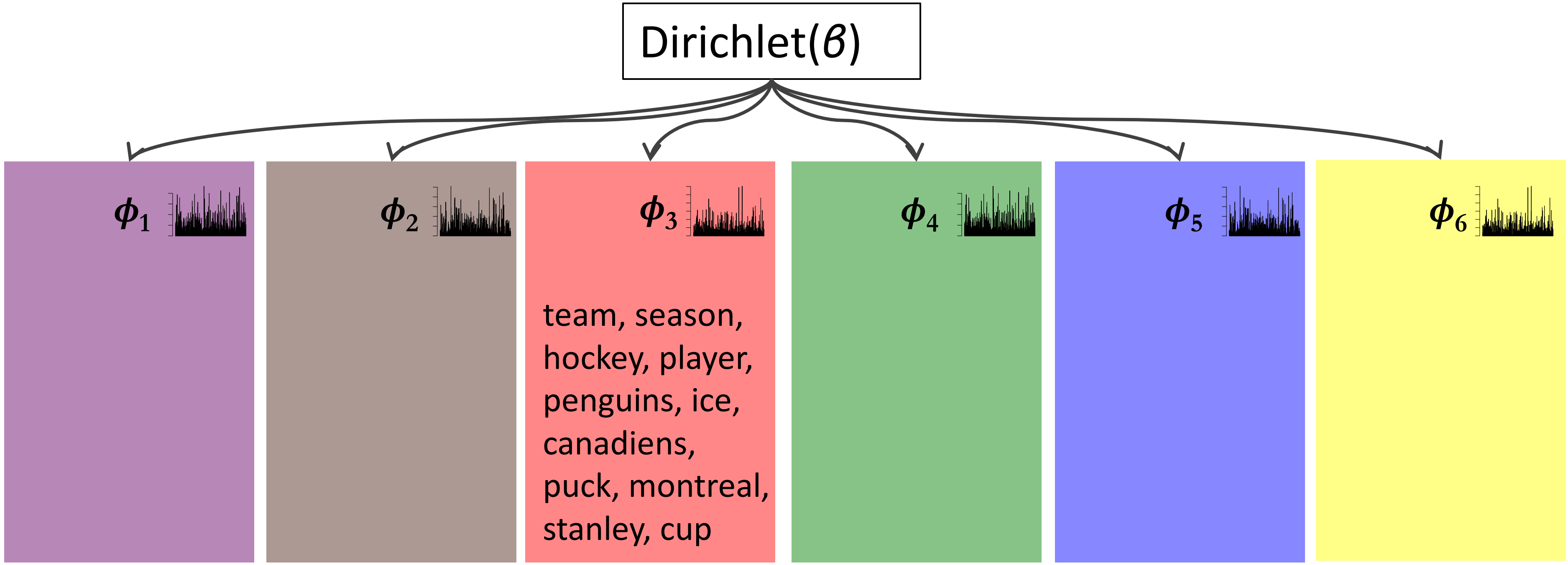
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by $\boldsymbol{\phi}_k$

(Blei, Ng, & Jordan, 2003)

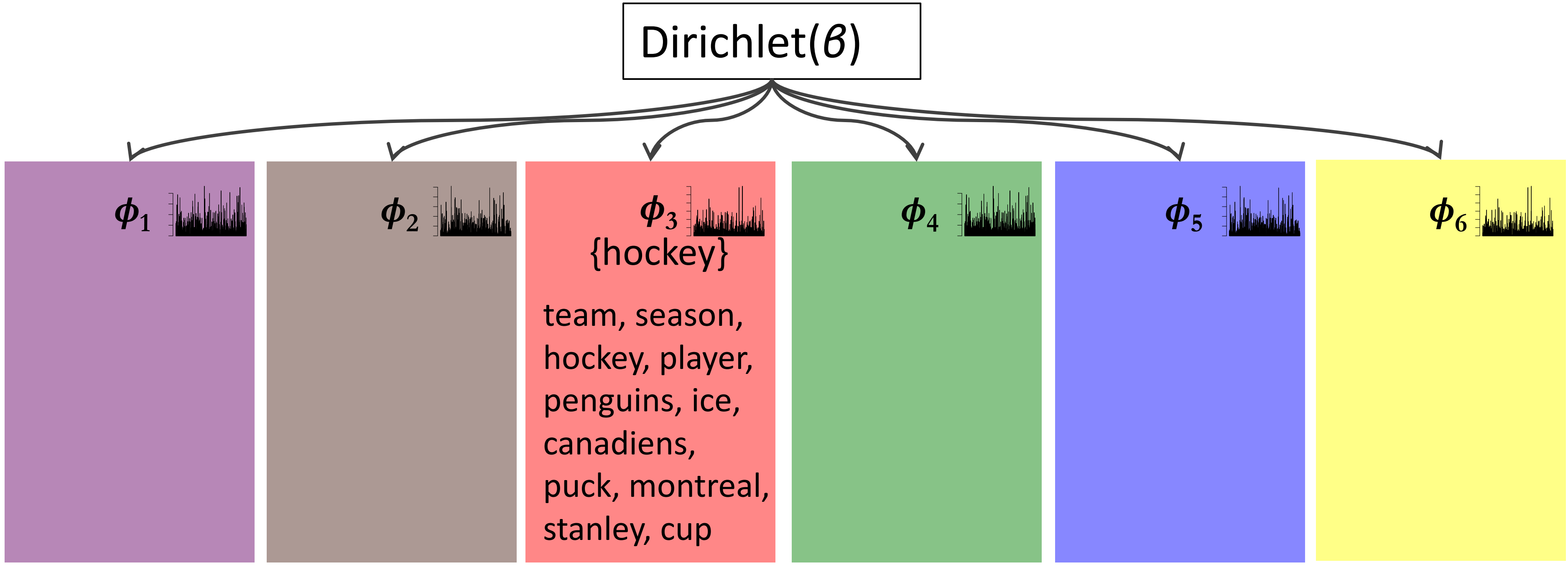
LDA for Topic Modeling



- A topic is visualized as its **high probability words.**

(Blei, Ng, & Jordan, 2003)

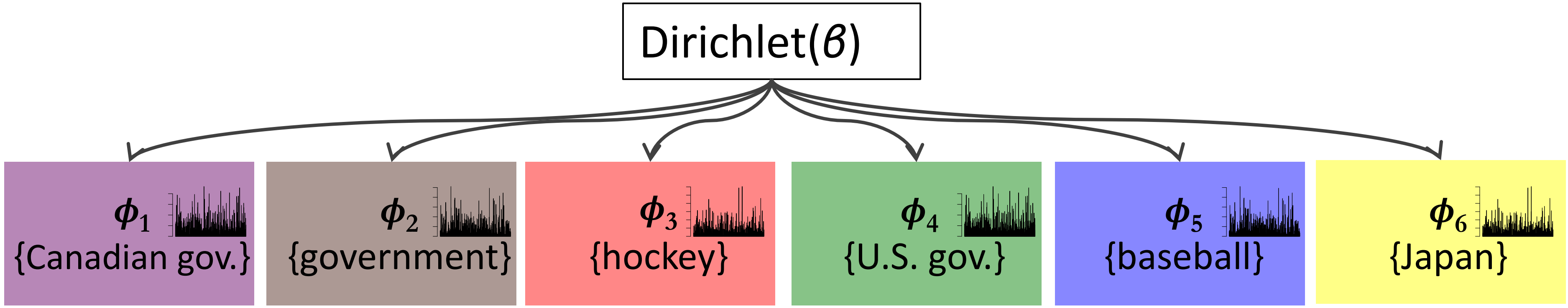
LDA for Topic Modeling



- A topic is visualized as its **high probability words.**
- A pedagogical **label** is used to identify the topic.

(Blei, Ng, & Jordan, 2003)

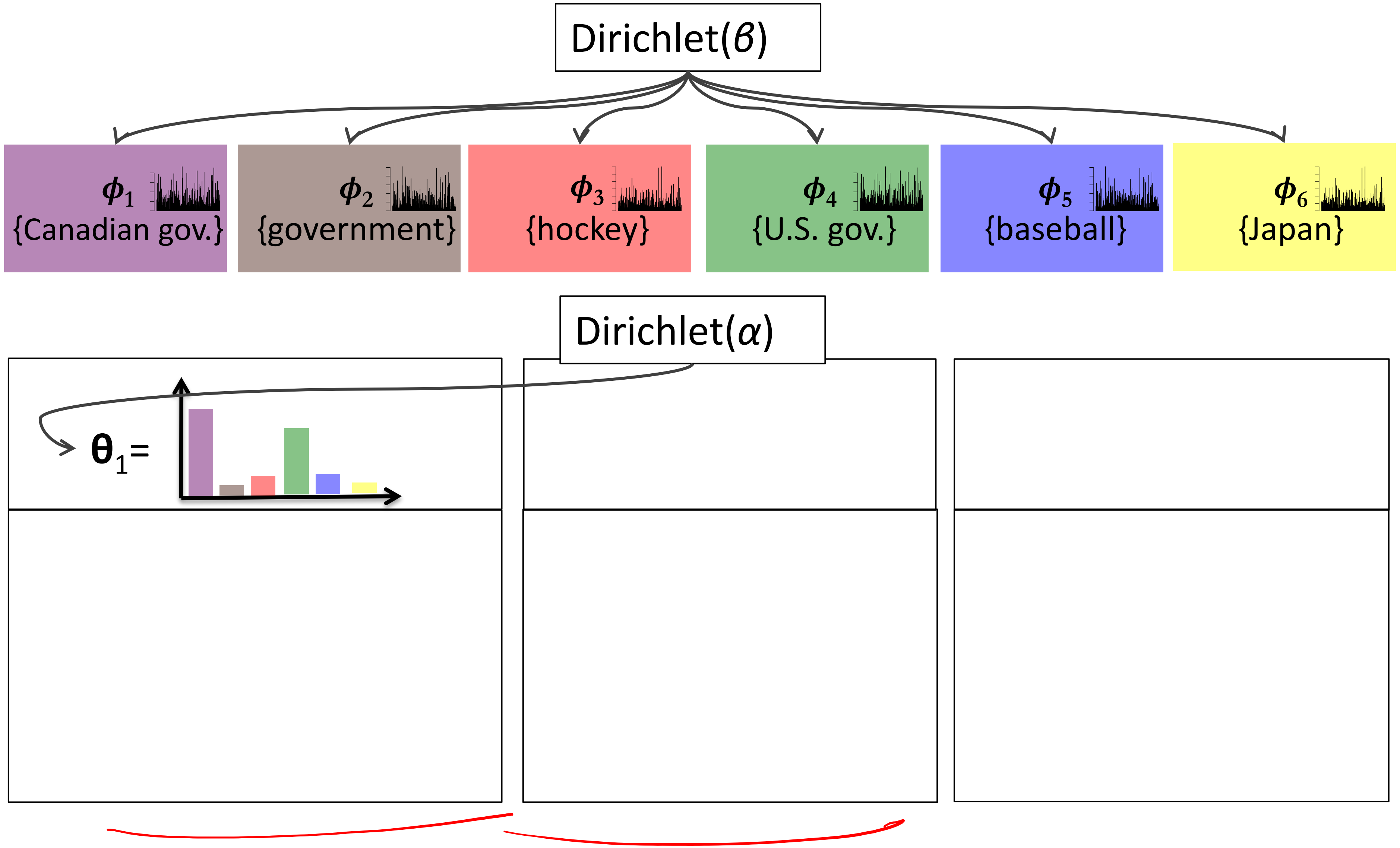
LDA for Topic Modeling



- A topic is visualized as its high probability words.
- A pedagogical **label** is used to identify the topic.

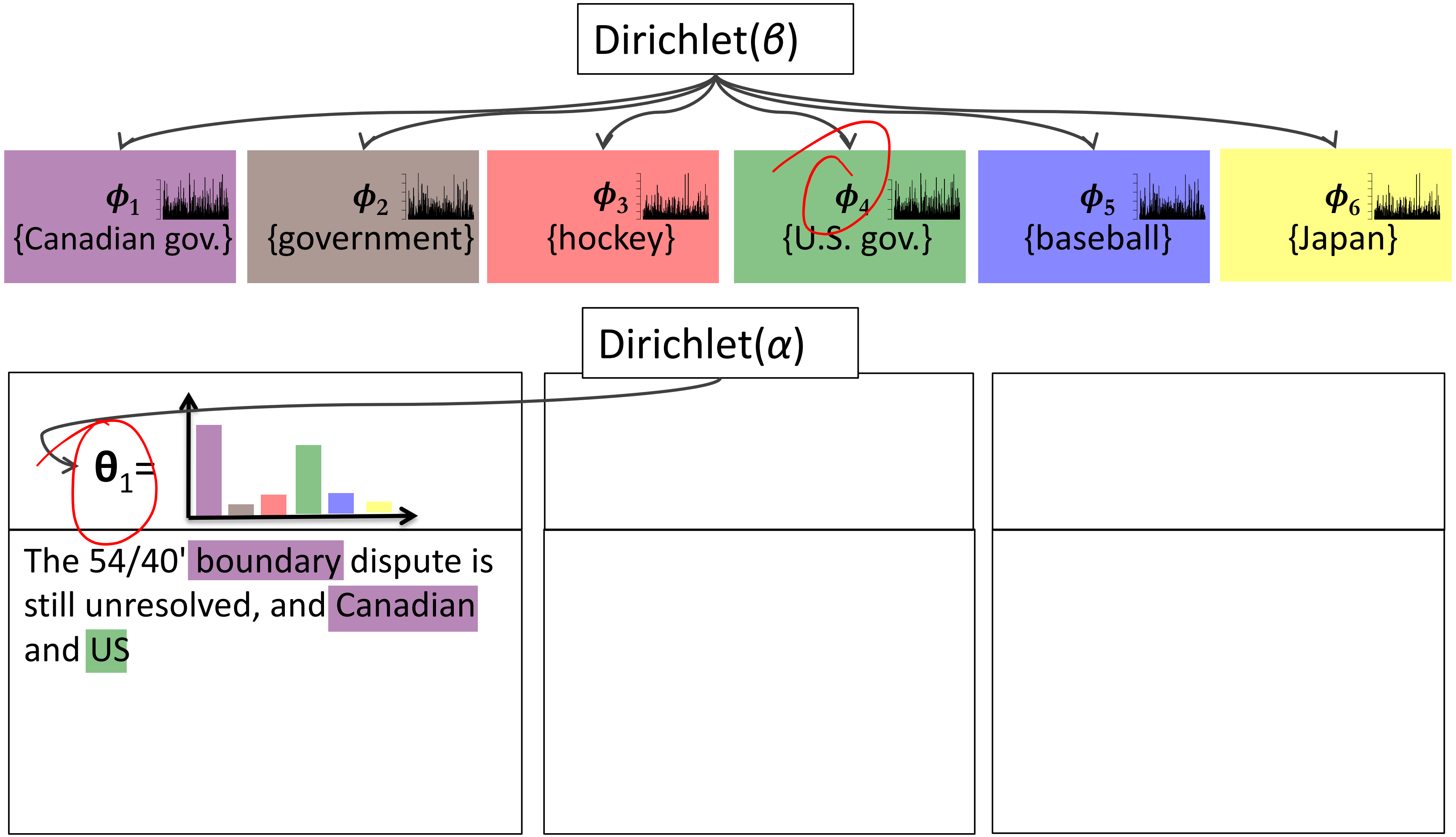
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



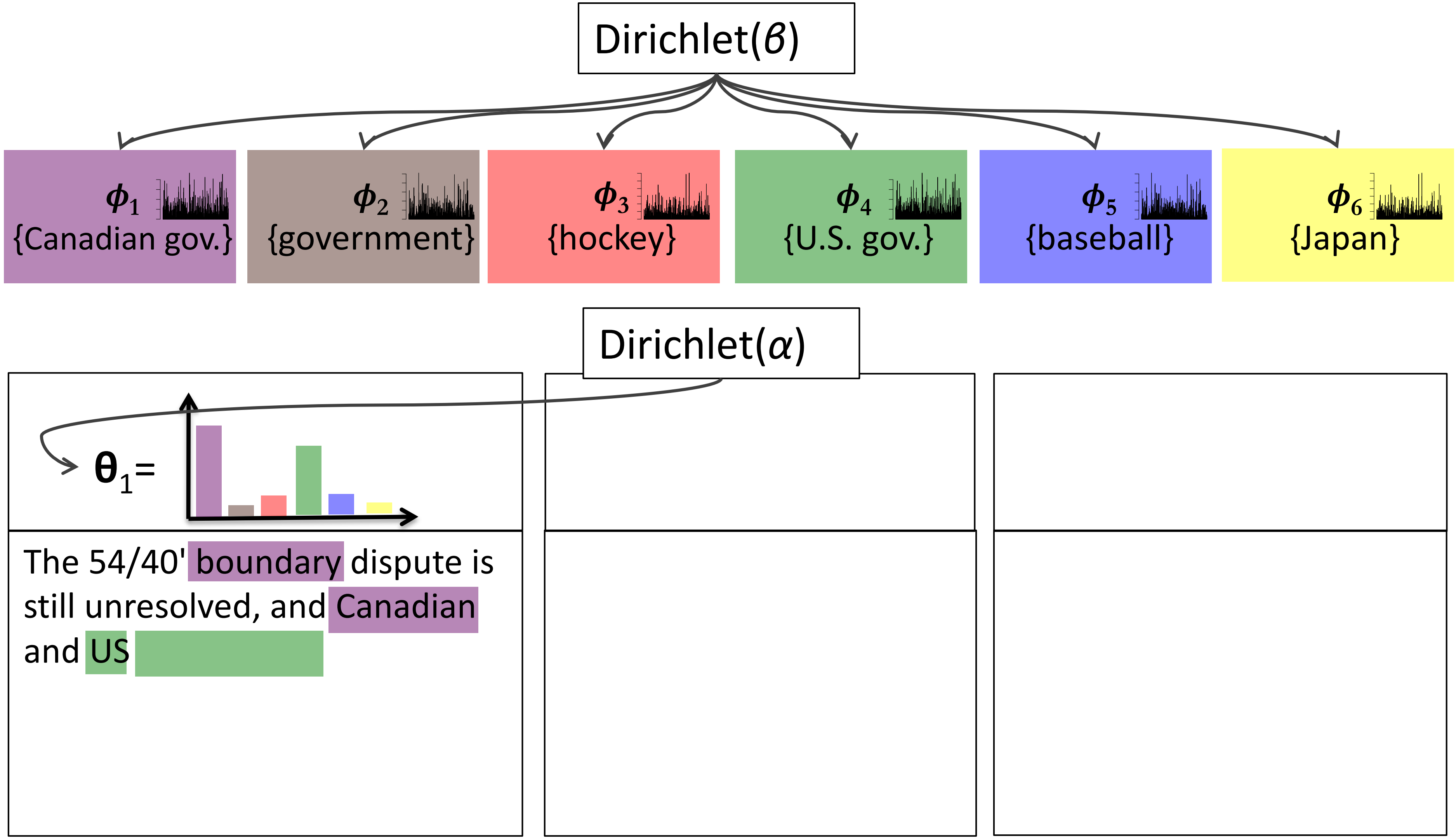
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



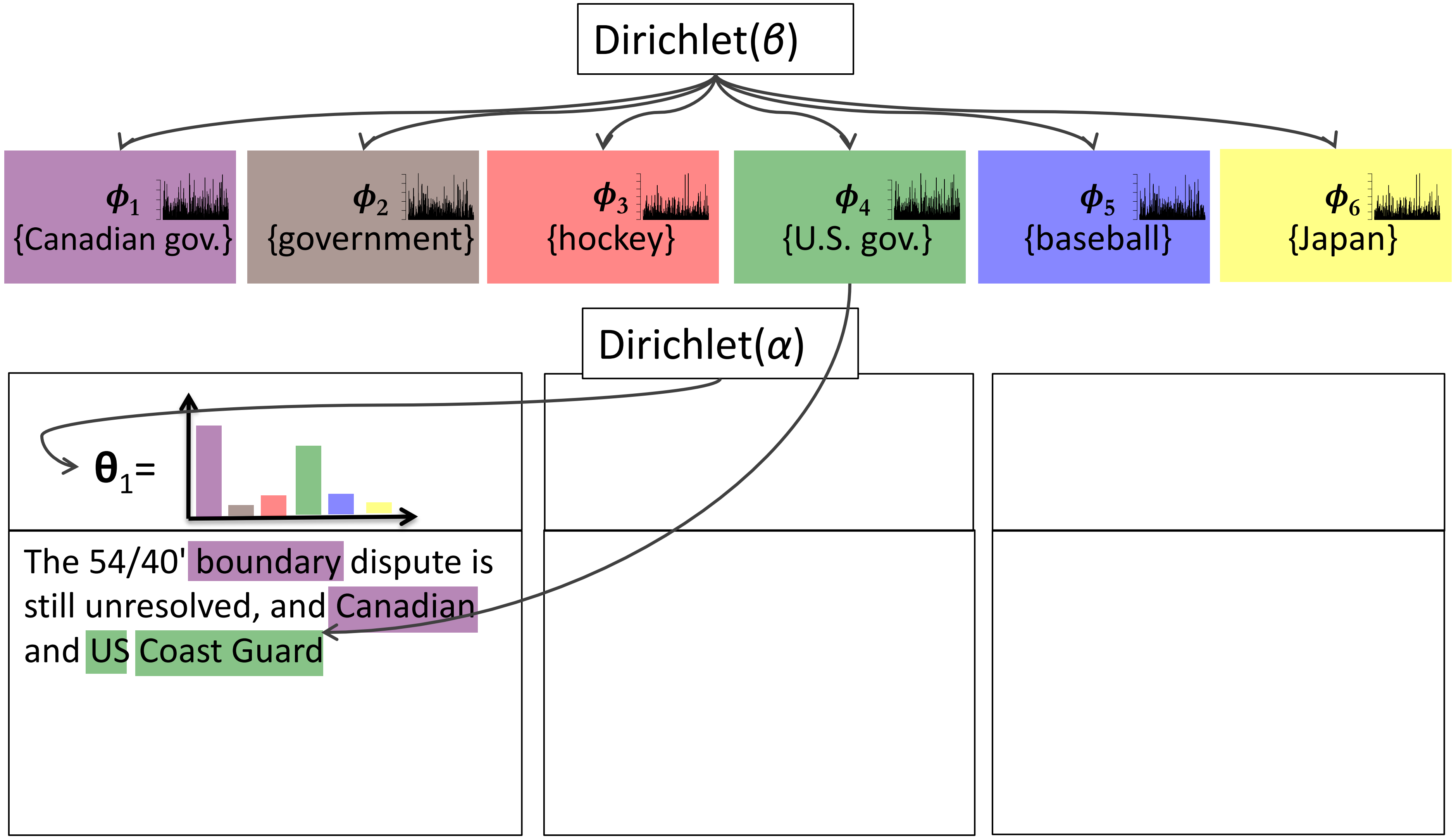
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



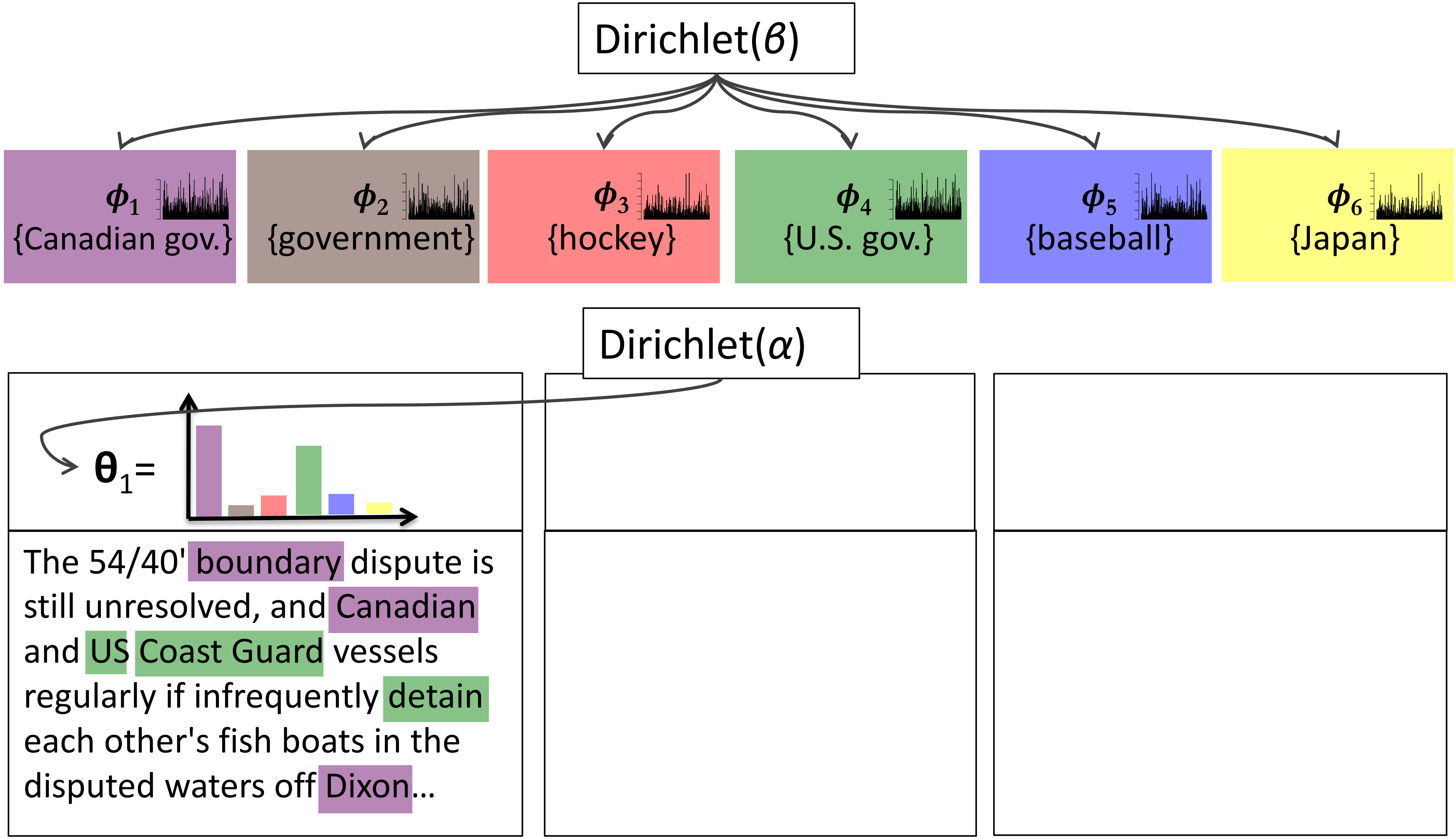
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



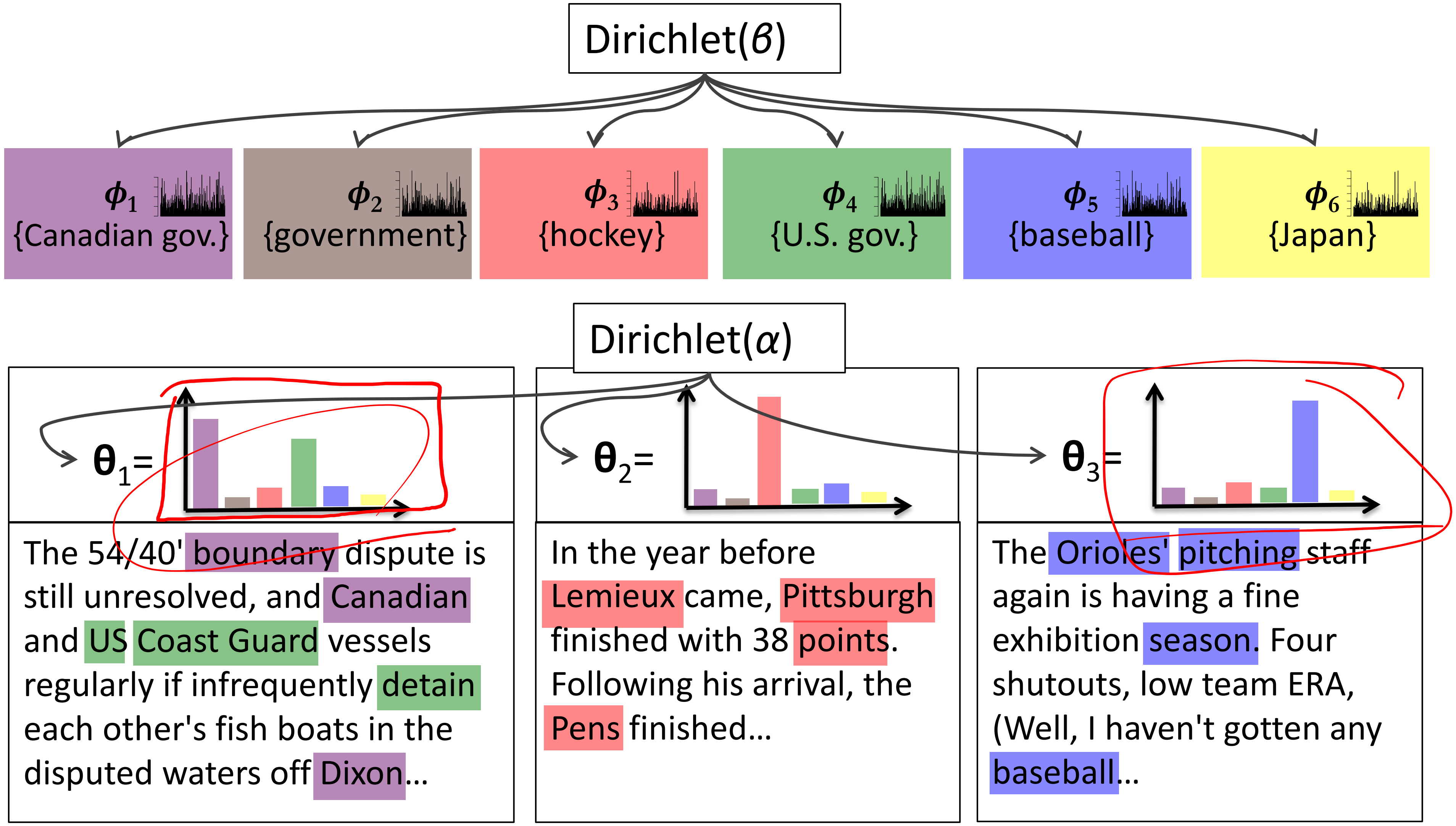
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



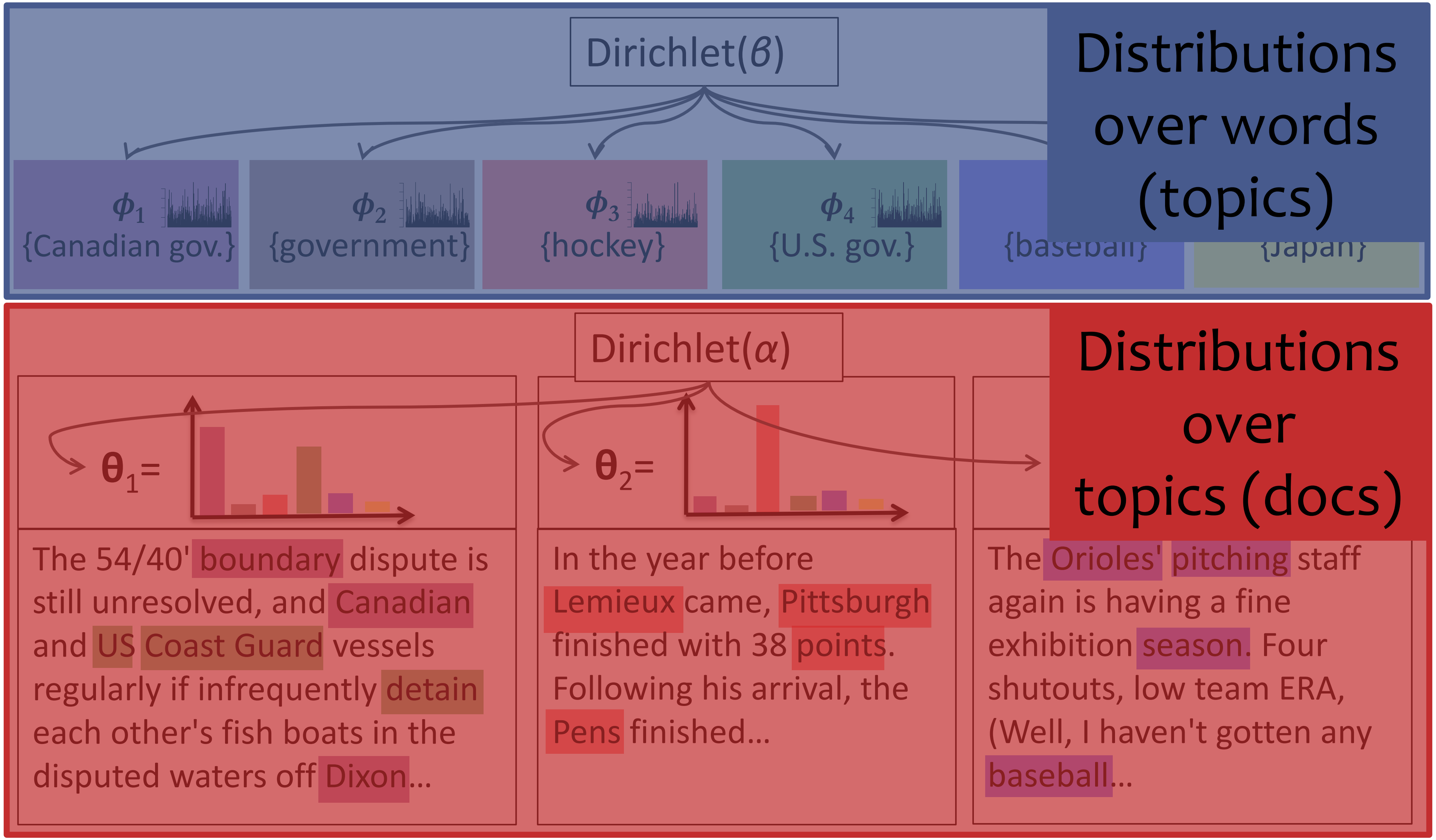
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



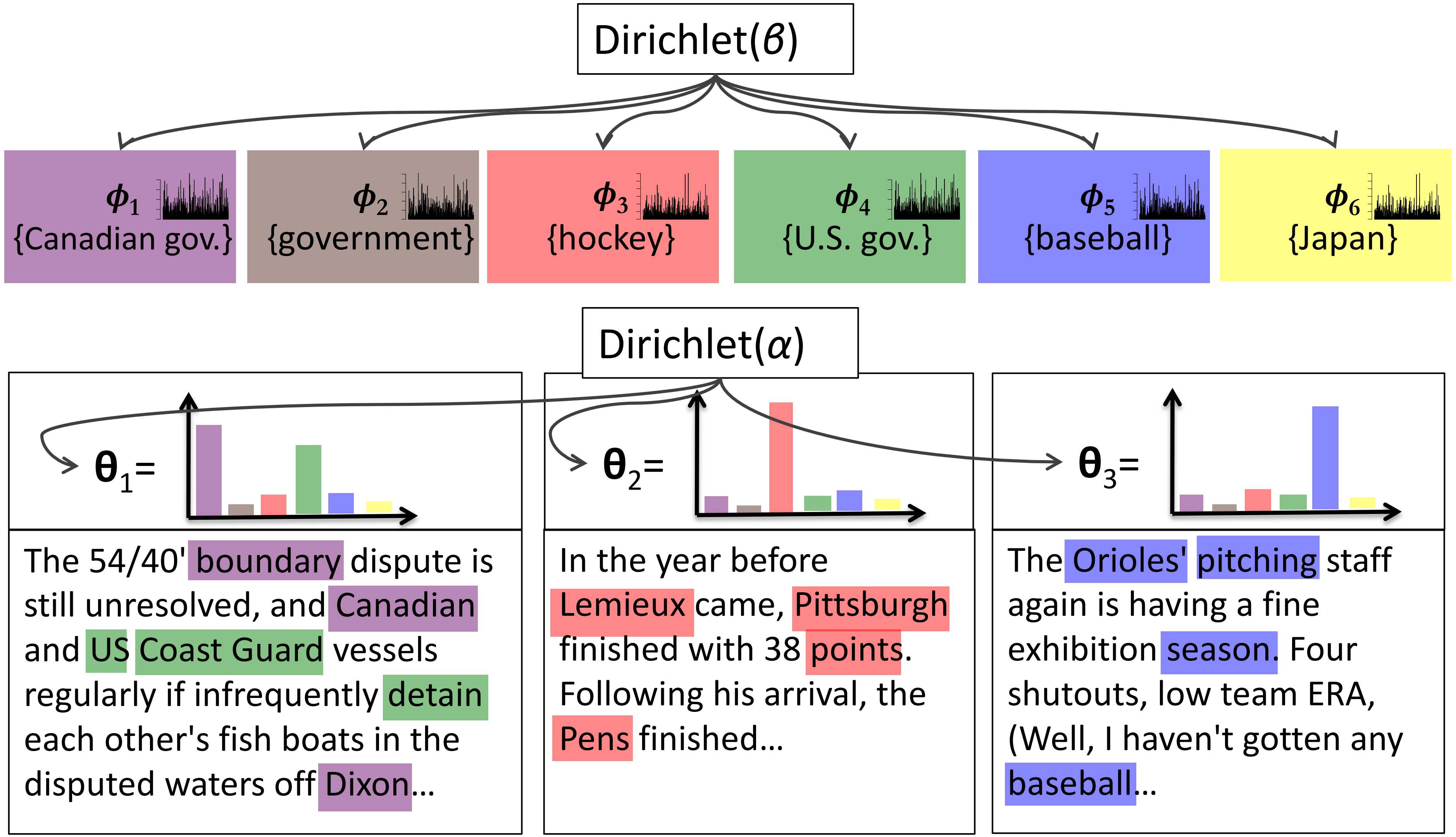
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



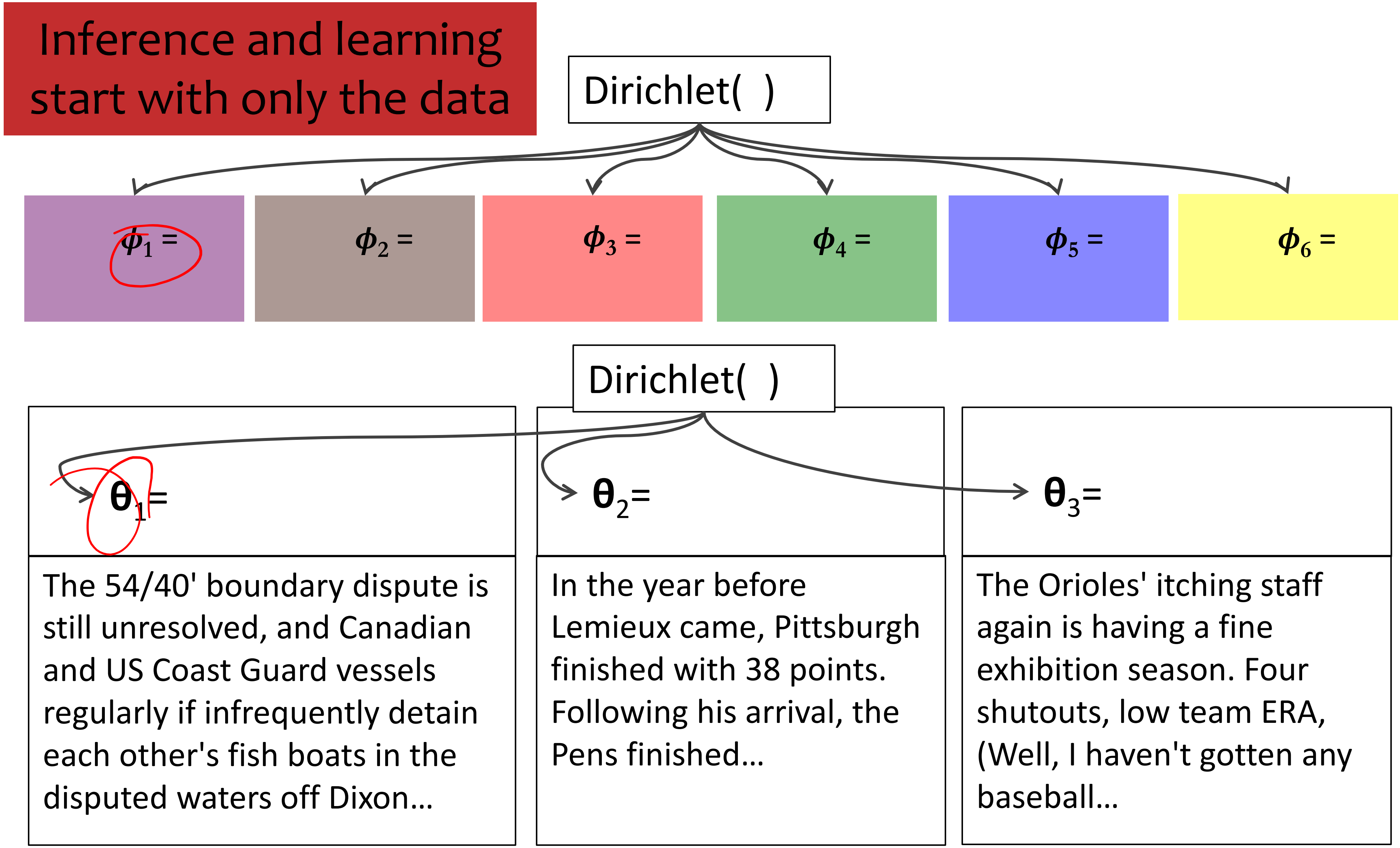
(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



(Blei, Ng, & Jordan, 2003)

LDA for Topic Modeling



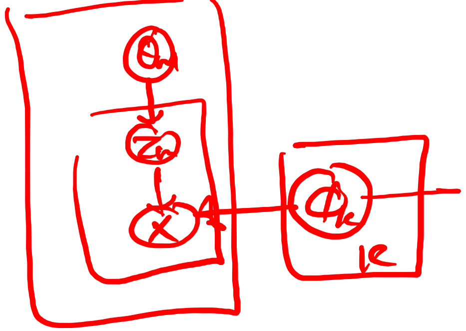
Latent Dirichlet Allocation

Questions:

- Is this a believable story for the generation of a corpus of documents?

- Why might it work well anyway?

Latent Dirichlet Allocation



Why does LDA “work”?

- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible. $\rightarrow \theta$
  2. For each topic, assign high probability to as few terms as possible. $\rightarrow \phi$
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document’s words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

Slide from David Blei, MLSS 2012

Latent Dirichlet Allocation

How does this relate to my other favorite model for capturing low-dimensional representations of a corpus?

- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999)
- It is a mixed-membership model (Erosheva, 2004).
- It relates to PCA and matrix factorization (Jakulin and Buntine, 2002)
- Was independently invented for genetics (Pritchard et al., 2000)

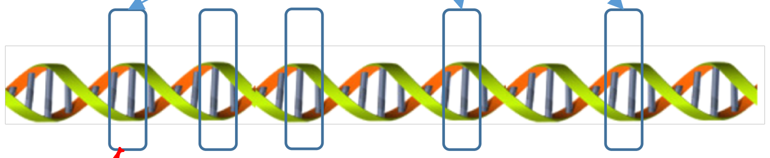
Slide from David Blei, MLSS 2012

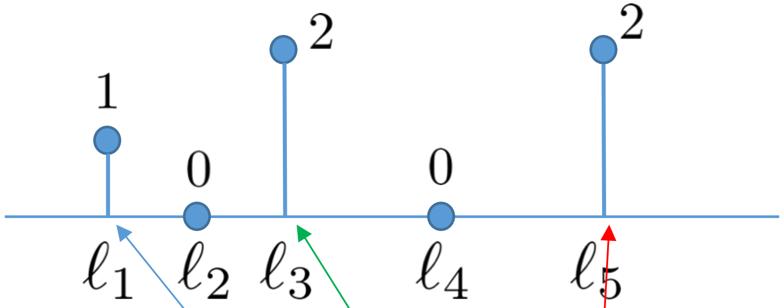
Case Study:
Modeling Join Imaging and Genetic data

Imaging and Genetic Data

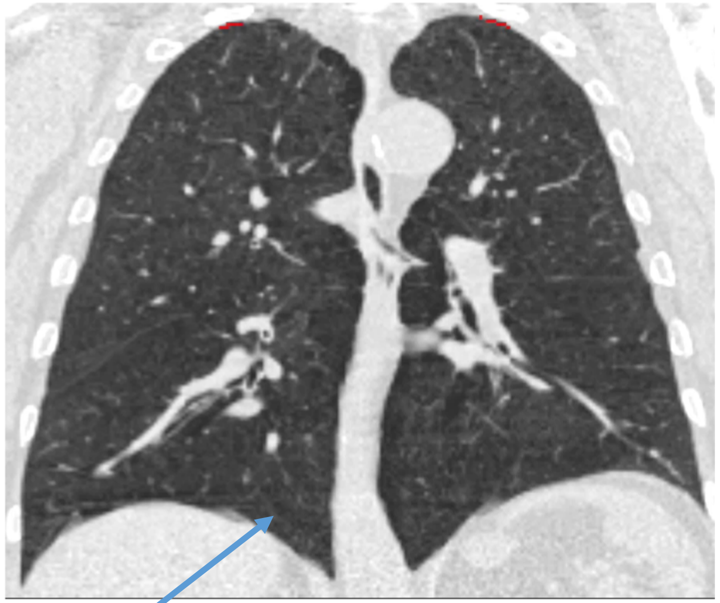
Subject s

Genetic loci of interest



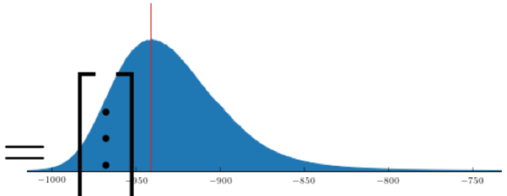


$$G_s = \{\ell_1, \ell_3, \ell_3, \ell_5, \ell_5\}$$



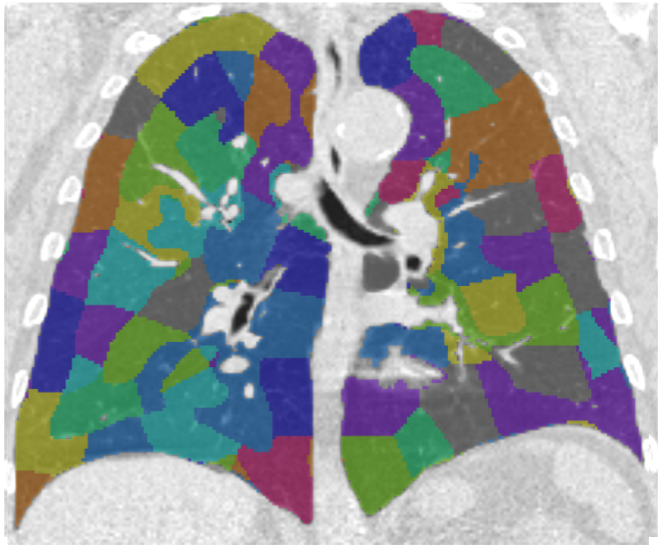
Imaging signature of Supervoxel n

$$I_{sn} = \begin{bmatrix} \vdots \end{bmatrix}$$



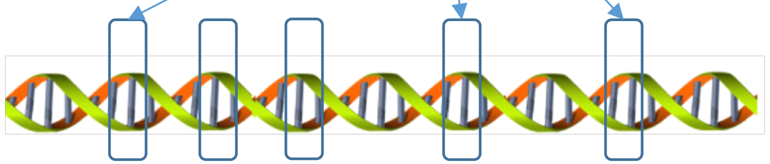
Bag of Words Model

Subject s



Visual Words (I_{sn})

Interesting Genetic loci

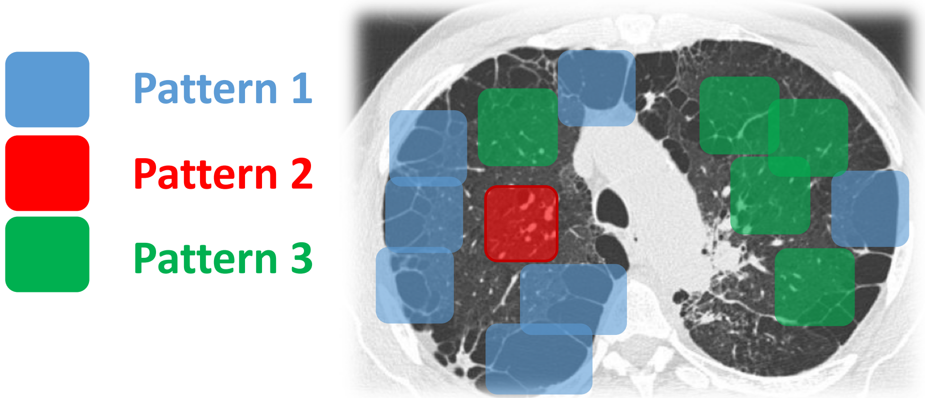


$$G_s = \{\ell_1, \ell_3, \ell_3, \ell_5, \ell_5\}$$

Genetic Words
(Genetic variants)

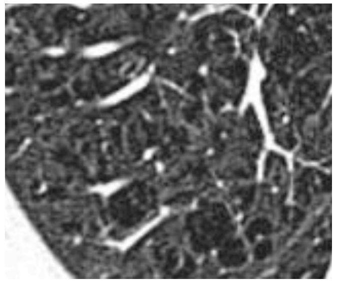
Subject ⟷ Document

Analogy: Subject as a Document

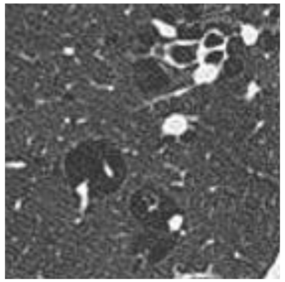


Topics (Image Patterns):

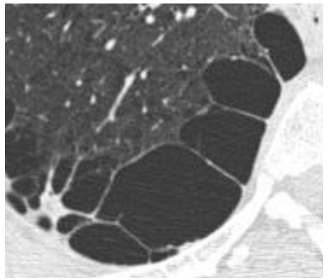
Pattern 1

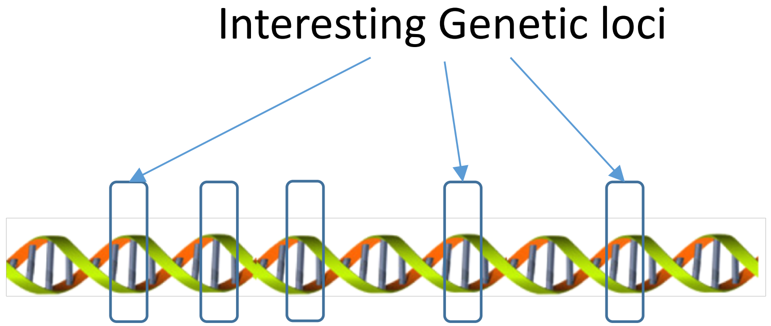


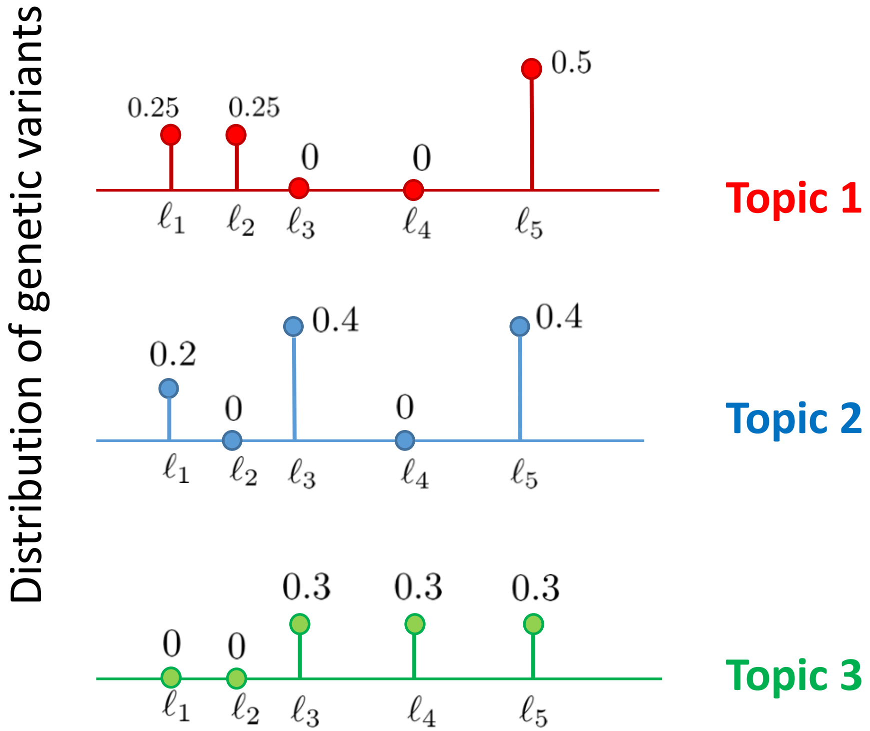
Pattern 2

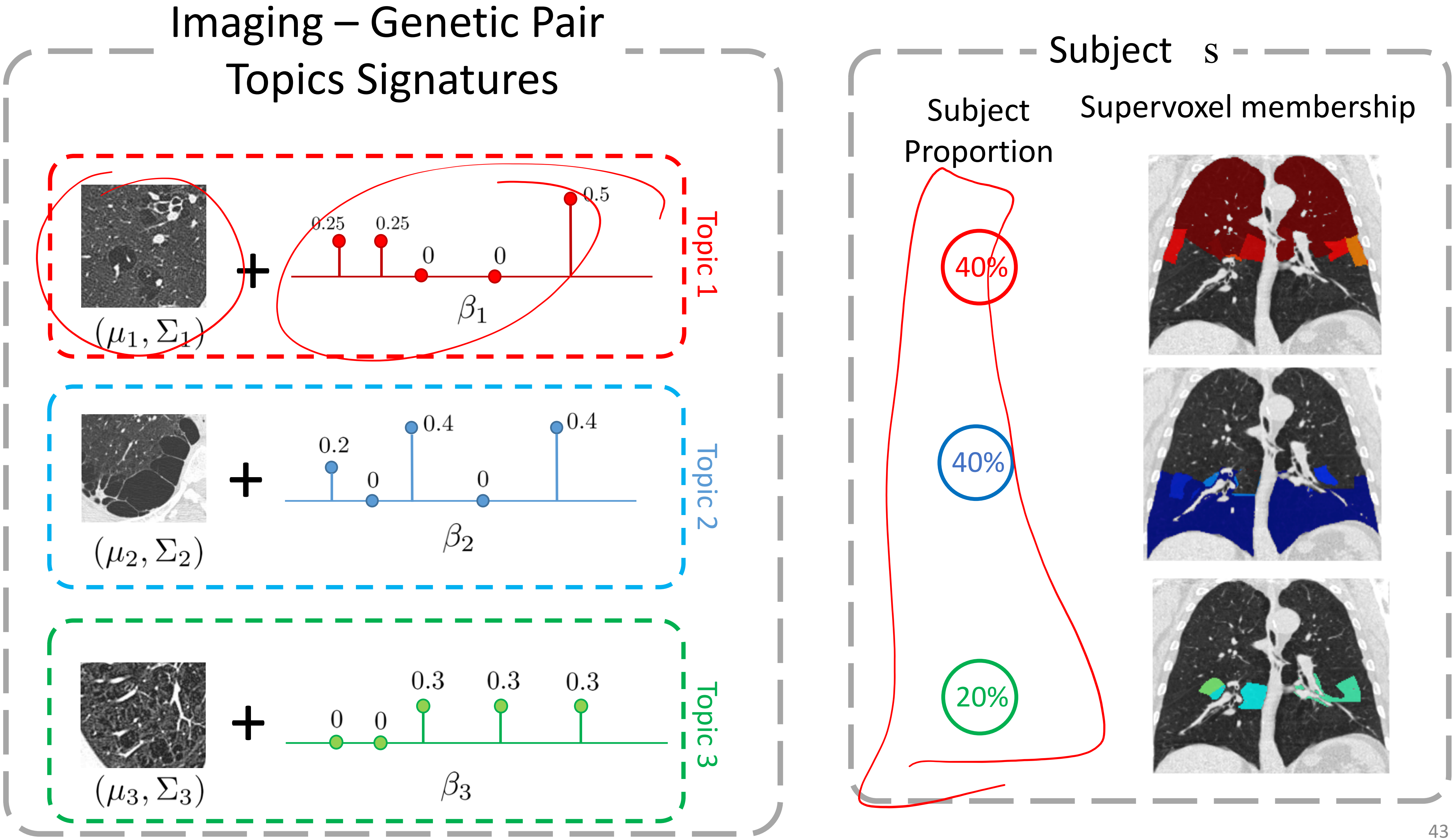


Pattern 3





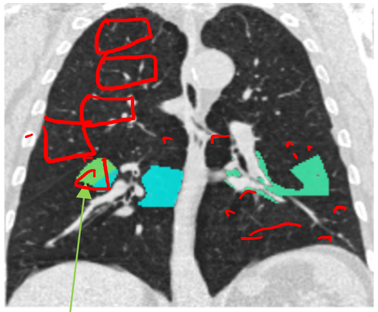


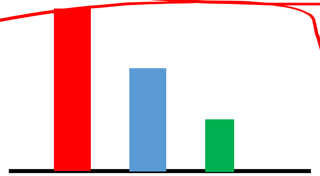

Imaging – Genetic Pair
Topics Signatures
0.25
0.25
0
0
0.5
(μ_1, Σ_1)
β_1
Topic 1
0.2
0
0.4
0
0.4
(μ_2, Σ_2)
β_2
Topic 2
0
0
0.3
0.3
0.3
(μ_3, Σ_3)
β_3
Topic 3
Subject s
Subject Proportion
Supervoxel membership
40%
40%
20%

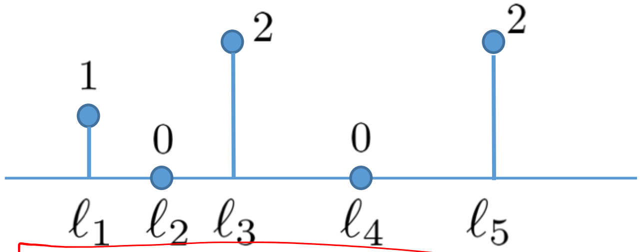
Probabilistic Model

$(\mu_1, \Sigma_1), \beta_1$ | $(\mu_2, \Sigma_2), \beta_2$ | $(\mu_3, \Sigma_3), \beta_3$ | $\cdots$ | $(\mu_K, \Sigma_K), \beta_K$

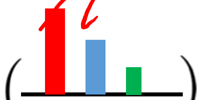
Subject s



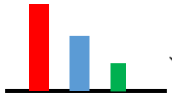




$$G_s = \{\ell_1, \ell_3, \boxed{\ell_3}, \ell_5, \ell_5\}$$

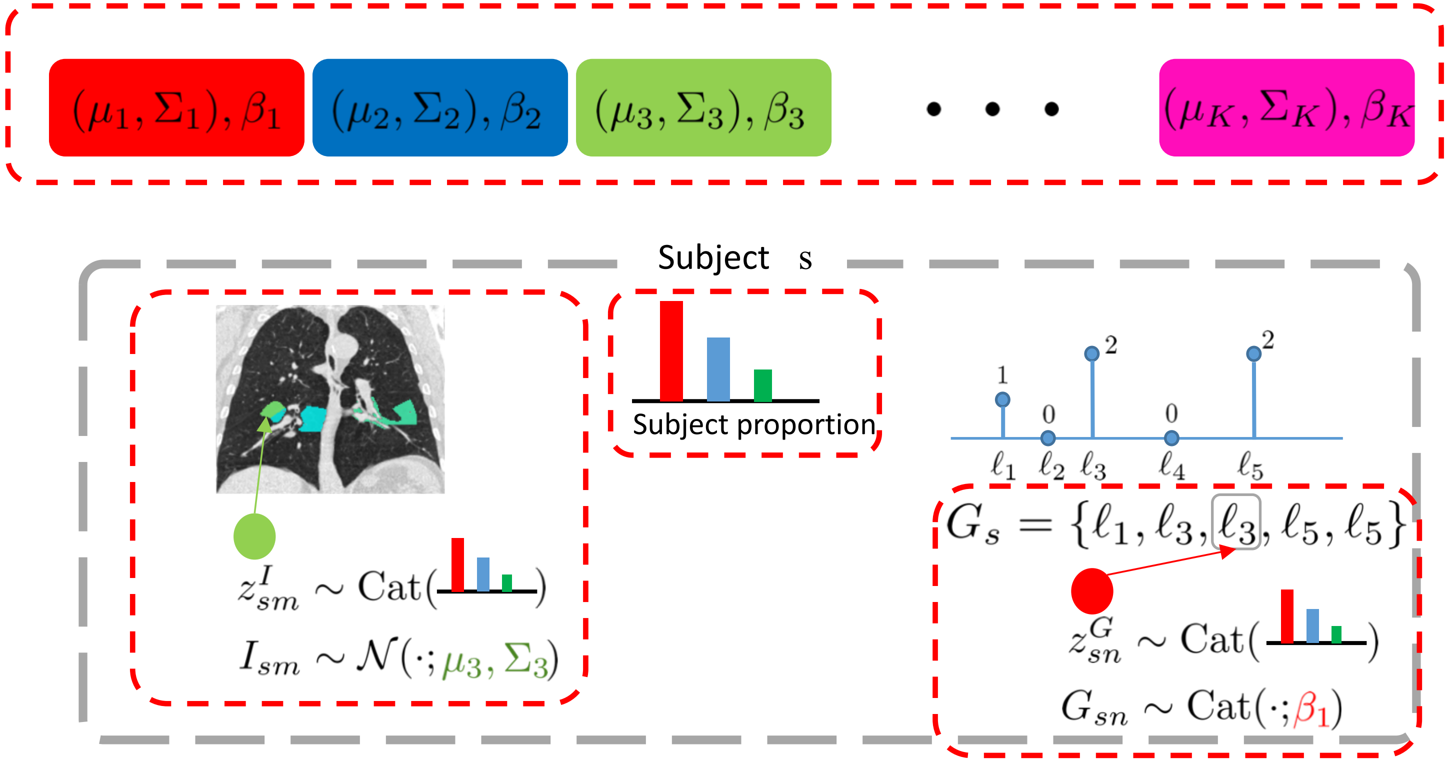
$z^I_{sm} \sim \mathrm{Cat}($  $)$

$$I_{sm} \sim \mathcal{N}(\cdot; \mu_3, \Sigma_3)$$

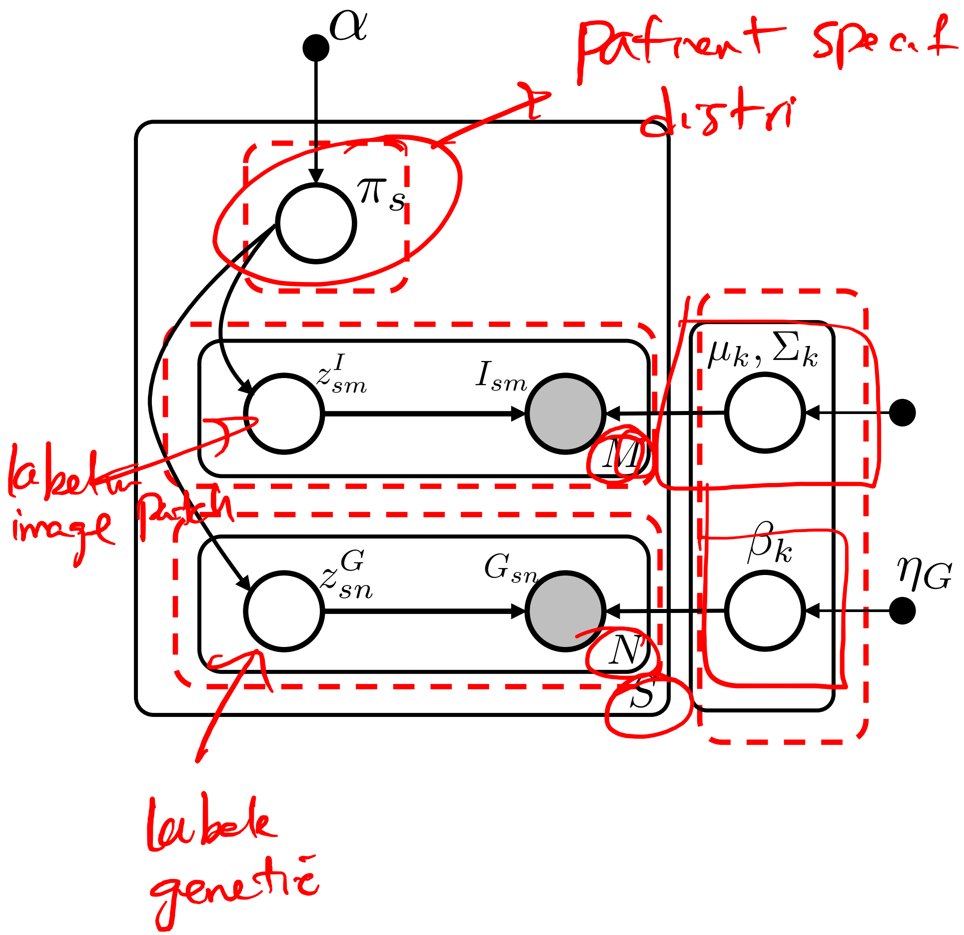
$z^G_{sn} \sim \mathrm{Cat}($  $)$

$$G_{sn} \sim \mathrm{Cat}(\cdot; \beta_1)$$

Graphical Model

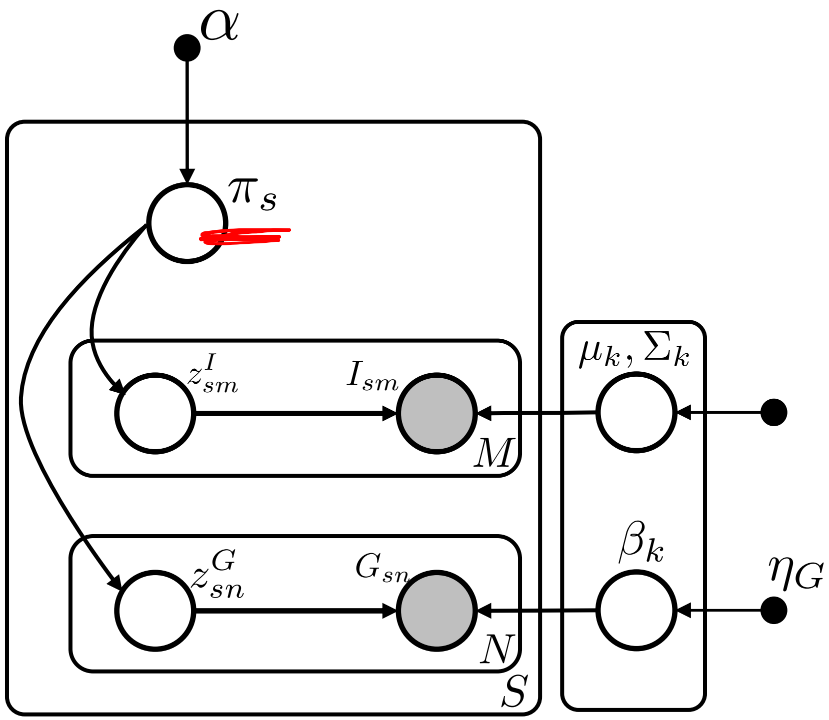


$$(\mu_k, \Sigma_k) \sim \mathrm{NIW}(\eta^I)$$
$$\beta_k \sim \mathrm{Dir}(\eta^G)$$



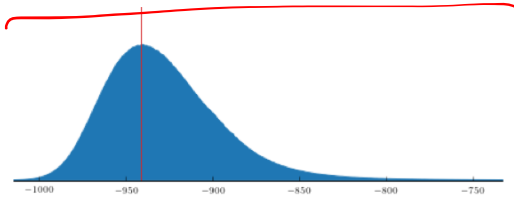
Inference

$(\mu_1, \Sigma_1), \beta_1$ | $(\mu_2, \Sigma_2), \beta_2$ | $(\mu_3, \Sigma_3), \beta_3$ | $\cdots$ | $(\mu_K, \Sigma_K), \beta_K$



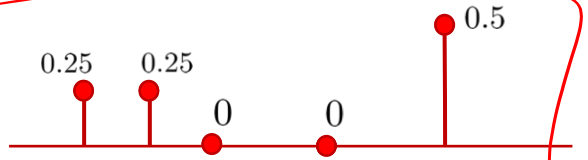
Topic pairs

$$p(\mu_k | \{I_{sm}\}, \{G_{sn}\}; \pi)$$



(Imaging signature)

$$p(\beta_k | \{I_{sm}\}, \{G_{sn}\}; \pi)$$

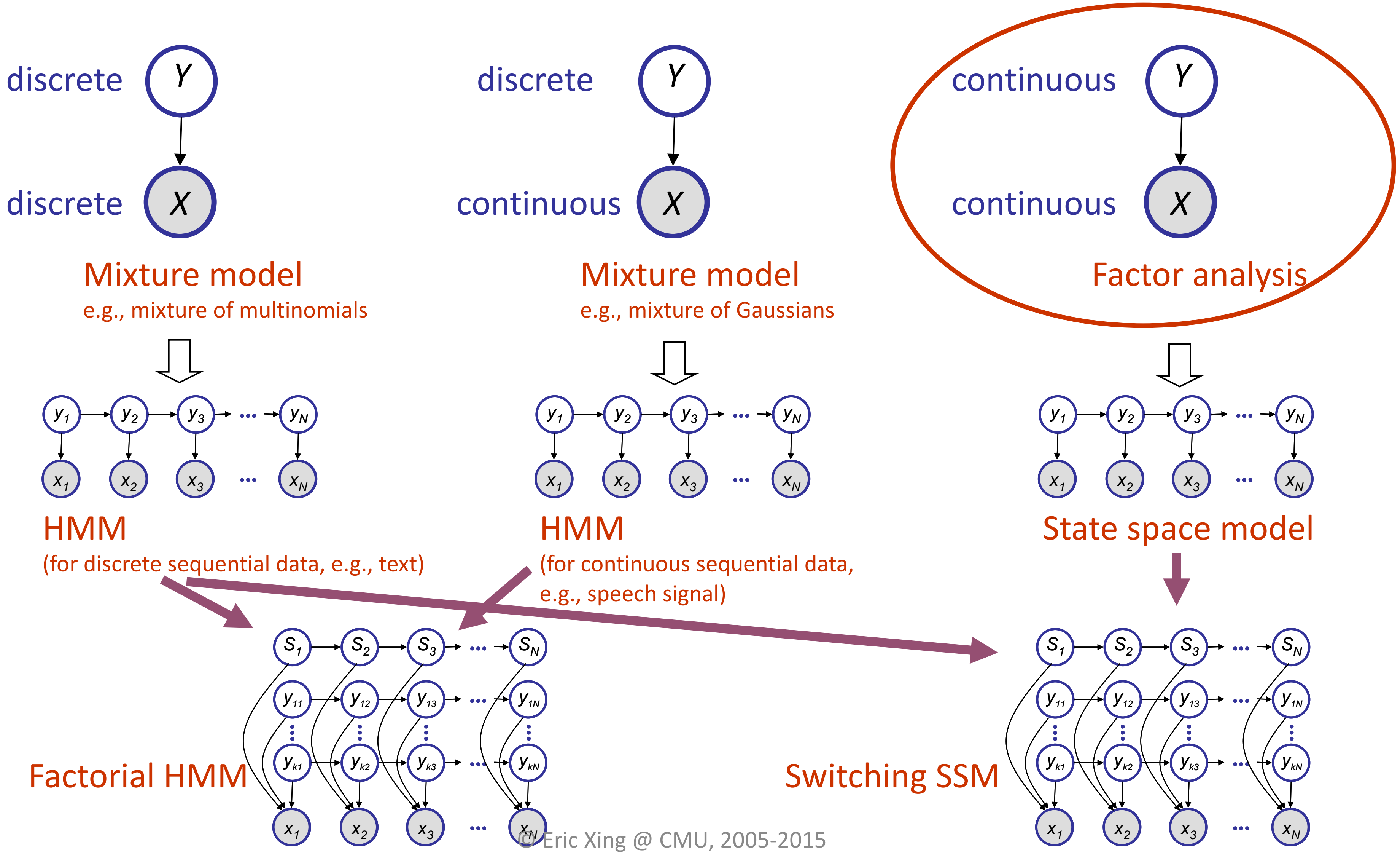


(Different Ranking SNPs)

$$\pi = \{\alpha, \omega, \eta^I, \eta^G\}$$ (hyper-parameters)

Factor Analysis

A road map to more complex dynamic models



Recall multivariate Gaussian

- Multivariate Gaussian density:

$$p(\mathbf{x}\mid\mu,\Sigma)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- A joint Gaussian:

$$p\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\mid\mu,\Sigma\right)=\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\middle|\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

- How to write down $p(\mathbf{x}_1)$, $p(\mathbf{x}_1|\mathbf{x}_2)$ or $p(\mathbf{x}_2|\mathbf{x}_1)$ using the block elements in μ and Σ?
  - Formulas to remember:

$$p(\mathbf{x}_2)=\mathcal{N}(\mathbf{x}_2\mid\mathbf{m}_2^m,\mathbf{V}_2^m)$$
$$\mathbf{m}_2^m=\mu_2$$
$$\mathbf{V}_2^m=\Sigma_{22}$$

$$p(\mathbf{x}_1|\mathbf{x}_2)=\mathcal{N}(\mathbf{x}_1\mid\mathbf{m}_{1|2},\mathbf{V}_{1|2})$$
$$\mathbf{m}_{1|2}=\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\mu_2)$$
$$\mathbf{V}_{1|2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Review: The matrix inverse lemma

- Consider a block-partitioned matrix: $M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$
- First we diagonalize M

$$\underbrace{\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}}_{X} \overbrace{\begin{bmatrix} E & F \\ G & H \end{bmatrix}}^{Y} \underbrace{\begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix}}_{Z} = \underbrace{\begin{bmatrix} E-FH^{-1}G & 0 \\ 0 & H \end{bmatrix}}_{W}$$

  - Schur complement: $M/H = E-FH^{-1}G$
- Then we inverse, using this formula: $XYZ = W \quad \Rightarrow \quad Y^{-1} = ZW^{-1}X$

$$M^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}$$

- Matrix inverse lemma

$$\left(E-FH^{-1}G\right)^{-1} = E^{-1} + E^{-1}F\left(H-GE^{-1}F\right)^{-1}GE^{-1}$$

Review: Some matrix algebra

$f(x) = ax$ $\qquad \frac{\partial f}{\partial} = a$

- Trace and derivatives $\qquad \mathrm{tr}[A] \stackrel{\mathrm{def}}{=} \sum_i a_{ii}$
  - Cyclical permutations

$$\mathrm{tr}[ABC] = \mathrm{tr}[CAB] = \mathrm{tr}[BCA]$$

  - Derivatives

$\langle B A \rangle$

$$\frac{\partial}{\partial A}\mathrm{tr}[BA] = B^T$$

$$\frac{\partial}{\partial A}\mathrm{tr}[x^T A x] = \frac{\partial}{\partial A}\mathrm{tr}[x x^T A] = x x^T$$

- Determinants and derivatives $\qquad \frac{\partial}{\partial A}\log|A| = A^{-1}$

Factor analysis

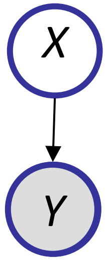
$p(y)$

$p(y|x)$

$y = \mu + \Lambda x + w$

$w \sim \mathcal{N}(0, \Psi)$

- An unsupervised linear regression model

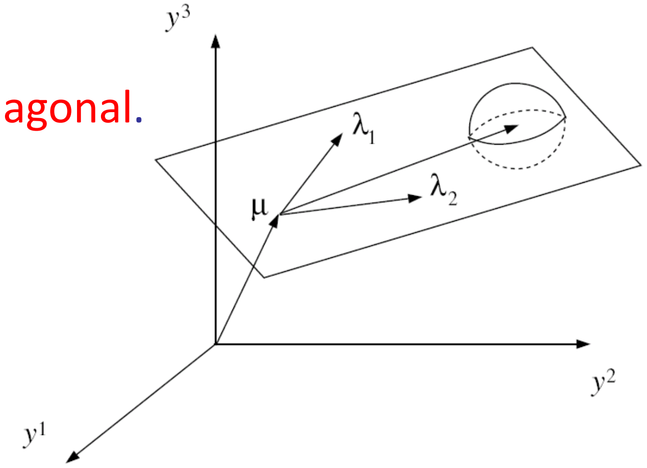


$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, I)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mu + \Lambda\mathbf{x}, \Psi)$$

where Λ is called a factor loading matrix, and Ψ is diagonal.

- Geometric interpretation

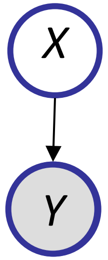


  - To generate data, first generate a point within the manifold then add noise. Coordinates of point are components of latent variable.

Marginal data distribution

- A marginal Gaussian (e.g., $p(\mathbf{x})$) times a conditional Gaussian (e.g., $p(\mathbf{y}|\mathbf{x})$) is a joint Gaussian
- Any marginal (e.g., $p(\mathbf{y})$ of a joint Gaussian (e.g., $p(\mathbf{x},\mathbf{y})$) is also a Gaussian
  - Since the marginal is Gaussian, we can determine it by just computing its mean and variance. (Assume noise uncorrelated with data.)

$$\begin{aligned} E[\mathbf{Y}] &= E[\mu + \Lambda\mathbf{X} + \mathbf{W}] \qquad \text{where } \mathbf{W} \sim \mathcal{N}(0, \Psi) \\ &= \mu + \Lambda E[\mathbf{X}] + E[\mathbf{W}] \\ &= \mu + 0 + 0 = \mu \end{aligned}$$



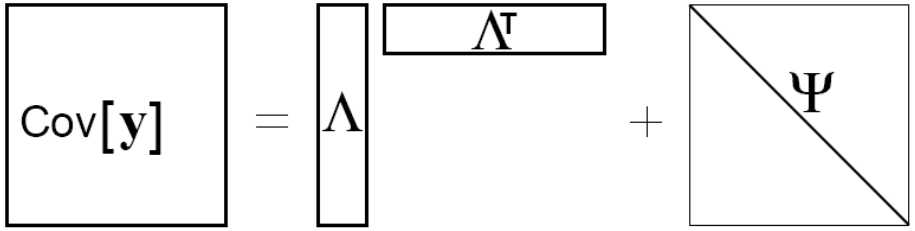
T

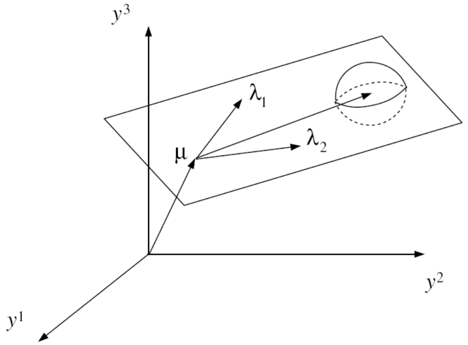
FA = Constrained-Covariance Gaussian

- Marginal density for factor analysis ($\mathbf{y}$ is p-dim, $\mathbf{x}$ is k-dim):

$$p(\mathbf{y}\,|\,\theta) = \mathcal{N}(\mathbf{y}; \mu, \Lambda\Lambda^T + \Psi)$$

- So the effective covariance is the low-rank outer product of two long skinny matrices plus a diagonal matrix:





- In other words, factor analysis is just a constrained Gaussian model (number of free params of the covariance is limited). (If Ψ were not diagonal then we could model any Gaussian and it would be pointless.)

FA joint distribution

- Model

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, I)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mu + \Lambda\mathbf{x}, \Psi)$$

- Covariance between $\mathbf{x}$ and $\mathbf{y}$

$$\begin{aligned} Cov[\mathbf{X}, \mathbf{Y}] &= E\left[(\mathbf{X} - 0)(\mathbf{Y} - \mu)^T\right] = E\left[\mathbf{X}(\mu + \Lambda\mathbf{X} + \mathbf{W} - \mu)^T\right] \\ &= E\left[\mathbf{X}\mathbf{X}^T\Lambda^T + \mathbf{X}\mathbf{W}^T\right] \\ &= \Lambda^T \end{aligned}$$

- Hence the joint distribution of $\mathbf{x}$ and $\mathbf{y}$:

$$p\left(\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix} \middle| \begin{bmatrix}0 \\ \mu\end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$$

- Assume noise is uncorrelated with data or latent variables.

$$\begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1}+H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix} = \begin{bmatrix} E^{-1}+E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}$$

$p(x,y)$ $\quad$ $p(x|y)$

Inference in Factor Analysis

- Apply the Gaussian conditioning formulas to the joint distribution we derived above, where

$$\Sigma_{11} = I$$
$$\Sigma_{12} = \Sigma_{12}^{\ T} = \Lambda^T$$
$$\Sigma_{22} = \left(\Lambda\Lambda^T + \Psi\right)$$

$p(y) \sim N(\underbrace{\Lambda}_{10\times 3} x + \mu, \Psi)$

$$\begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T+\Psi \end{bmatrix}$$

we can now derive the posterior of the latent variable $\mathbf{x}$ given observation $\mathbf{y}$, $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x} \mid \mathbf{m}_{1|2}, \mathbf{V}_{1|2})$, where

$$\mathbf{m}_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \mu_2) \qquad \mathbf{V}_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$
$$= \Lambda^T\left(\Lambda\Lambda^T + \Psi\right)^{-1}(\mathbf{y} - \mu) \qquad = I - \Lambda^T\left(\underbrace{\Lambda\Lambda^T}_{10\times 10} + \Psi\right)^{-1}\Lambda$$

Applying the matrix inversion lemma

$$\left(E\text{-}FH^{-1}G\right)^{-1} = E^{-1} + E^{-1}F\left(H\text{-}GE^{-1}F\right)^{-1}GE^{-1}$$

$$\Rightarrow \quad \mathbf{V}_{1|2} = \underbrace{\left(I + \Lambda^T\Psi^{-1}\Lambda\right)^{-1}}_{3\times 3} \qquad \mathbf{m}_{1|2} = \mathbf{V}_{1|2}\Lambda^T\Psi^{-1}(\mathbf{y} - \mu)$$

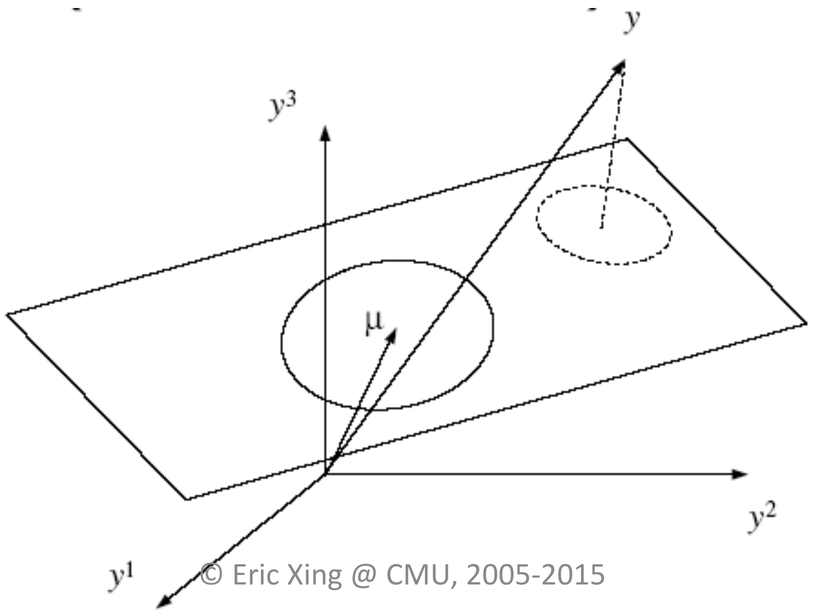
- Here we only need to invert a matrix of size $|\mathbf{x}|\times|\mathbf{x}|$, instead of $|\mathbf{y}|\times|\mathbf{y}|$.

Geometric interpretation: inference is linear projection

- The posterior is: $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \mathbf{m}_{1|2}, \mathbf{V}_{1|2})$

$$\mathbf{m}_{1|2} = \mathbf{V}_{1|2}\Lambda^T\Psi^{-1}(\mathbf{y}-\mu) \qquad \mathbf{V}_{1|2} = \left(I + \Lambda^T\Psi^{-1}\Lambda\right)^{-1}$$

- Posterior covariance does not depend on observed data $\mathbf{y}$!
- Computing the posterior mean is just a linear operation:



Learning FA

- Now, assume that we are given $\{y_n\}$ (the observation on high-dimensional data) only

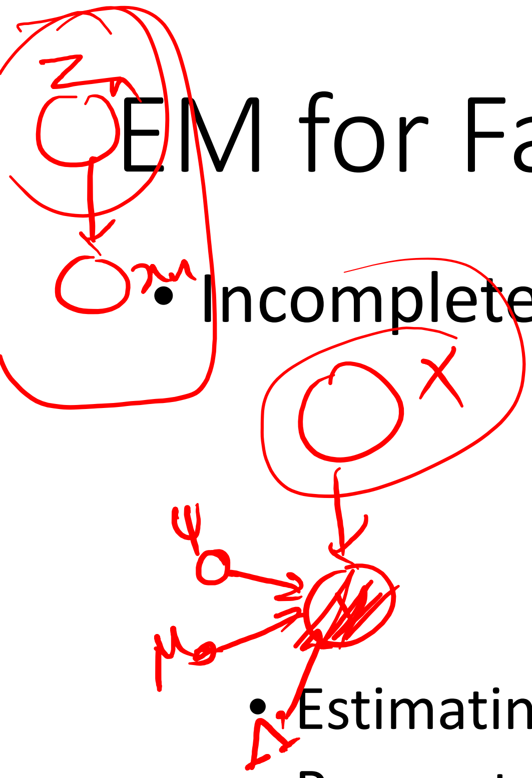
- We have derived how to estimate x_n from $P(X|Y)$

- How can we learning the model?
  - Loading matrix Λ
  - Manifold center μ
  - Variance Ψ

EM for Factor Analysis

$P(x_i;\theta) = \sum \pi_k N(x_i; \mu_k, \Sigma_k)$

$\theta = \{\mu_k, \Sigma_k, \pi_k$



- Incomplete data log likelihood function (marginal density of y)

$$\ell(\theta, D) = -\frac{N}{2}\log\left|\Lambda\Lambda^T + \Psi\right| - \frac{1}{2}\sum_n \boxed{(y_n - \mu)^T \left(\Lambda\Lambda^T + \Psi\right)^{-1}(y_n - \mu)} \rightarrow \mathrm{trace}\left((\Lambda\Lambda^T + \Psi)^{-1}\right.$$

$$= -\frac{N}{2}\log\left|\Lambda\Lambda^T + \Psi\right| - \frac{1}{2}\mathrm{tr}\left[\left(\Lambda\Lambda^T + \Psi\right)^{-1}\mathbf{S}\right], \qquad \text{where } \boxed{\mathbf{S} = \sum_n (y_n - \mu)(y_n - \mu)^T}$$

$P(X,Y)$

$P(Y)$

  - Estimating μ is trivial: $\hat{\mu}^{ML} = \frac{1}{N}\sum_n y_n$
  - Parameters Λ and Ψ are coupled nonlinearly in log-likelihood

- Complete log likelihood

$\langle \; \rangle$

$P(X|D)$

$$\ell_c(\theta, D) = \sum_n \log p(x_n, y_n) = \sum_n \log p(x_n) + \log p(y_n \mid x_n)$$

$$= -\frac{N}{2}\log|I| - \frac{1}{2}\sum_n x_n^T x_n - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n (y_n - \Lambda x_n)^T \Psi^{-1}(y_n - \Lambda x_n)$$

$$= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[x_n x_n^T\right] - \frac{N}{2}\mathrm{tr}\left[\mathbf{S}\Psi^{-1}\right], \qquad \text{where } \boxed{\mathbf{S} = \frac{1}{N}\sum_n (y_n - \Lambda x_n)(y_n - \Lambda x_n)^T}$$

unobs

Data

E-step for Factor Analysis

- Compute $\langle \ell_c(\theta, D) \rangle_{p(x|y)}$

$$\langle \ell_c(\theta, D) \rangle = -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[\langle X_n X_n^T \rangle\right] - \frac{N}{2}\mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right]$$

$$\langle \mathbf{S} \rangle = \frac{1}{N}\sum_n (y_n y_n^T - y_n \langle X_n^T \rangle \Lambda^T - \Lambda \langle X_n^T \rangle y_n^T + \Lambda \langle X_n X_n^T \rangle \Lambda^T)$$

$$\langle X_n \rangle = E[X_n \mid y_n]$$

$$\langle X_n X_n^T \rangle = Var[X_n \mid y_n] + E[X_n \mid y_n] E[X_n \mid y_n]^T$$

  - Recall that we have derived:

$$\mathbf{V}_{1|2} = \left(I + \Lambda^T \Psi^{-1} \Lambda\right)^{-1} \qquad \mathbf{m}_{1|2} = \mathbf{V}_{1|2} \Lambda^T \Psi^{-1} (\mathbf{y} - \mu)$$

$$\Rightarrow \quad \langle X_n \rangle = \mathbf{m}_{x_n|y_n} = \mathbf{V}_{1|2} \Lambda^T \Psi^{-1} (y_n - \mu) \quad \text{and} \quad \langle X_n X_n^T \rangle = \mathbf{V}_{1|2} + \mathbf{m}_{x_n|y_n} \mathbf{m}_{x_n|y_n}^T$$

Λ

M-step for Factor Analysis

$Y = X\beta + \varepsilon$

$\beta = (X^T X)^{-1} X^T y$

- Take the derivates of the expected complete log likelihood wrt. parameters.
  - Using the trace and determinant derivative rules:

$$\frac{\partial}{\partial \Psi^{-1}} \langle \ell_c \rangle = \frac{\partial}{\partial \Psi^{-1}} \left(-\frac{N}{2} \log|\Psi| - \frac{1}{2} \sum_n \mathrm{tr}\left[\langle X_n X_n^T \rangle\right] - \frac{N}{2} \mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right] \right)$$

$$= \frac{N}{2} \Psi - \frac{N}{2} \langle \mathbf{S} \rangle \qquad \Rightarrow \qquad \boxed{\Psi^{t+1} = \langle \mathbf{S} \rangle}$$

$$\frac{\partial}{\partial \Lambda} \langle \ell_c \rangle = \frac{\partial}{\partial \Lambda} \left(-\frac{N}{2} \log|\Psi| - \frac{1}{2} \sum_n \mathrm{tr}\left[\langle X_n X_n^T \rangle\right] - \frac{N}{2} \mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right] \right) = -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \Lambda} \langle \mathbf{S} \rangle$$

$$= -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \Lambda} \left(\frac{1}{N} \sum_n \left(y_n y_n^T - y_n \langle X_n^T \rangle \Lambda^T - \Lambda \langle X_n^T \rangle y_n^T + \Lambda \langle X_n X_n^T \rangle \Lambda^T \right) \right)$$

$$= \Psi^{-1} \sum_n y_n \langle X_n^T \rangle - \Psi^{-1} \Lambda \sum_n \langle X_n X_n^T \rangle \qquad \Rightarrow \qquad \boxed{\Lambda^{t+1} = \left(\sum_n y_n \langle X_n^T \rangle \right) \left(\sum_n \langle X_n X_n^T \rangle \right)^{-1}}$$

β

E

$y|x$ $y = \Lambda x + \mu$

Model Invariance and Identifiability

- There is *degeneracy* in the FA model.
- Since Λ only appears as outer product $\Lambda\Lambda^T$, the model is invariant to rotation and axis flips of the latent space.
- We can replace Λ with ΛQ for any orthonormal matrix Q and the model remains the same: $(\Lambda Q)(\Lambda Q)^T = \Lambda(QQ^T)\Lambda^T = \Lambda\Lambda^T$.
- This means that there is no "one best" setting of the parameters. An infinite number of parameters all give the ML score!
- Such models are called un-identifiable since two people both fitting ML parameters to the identical data will not be guaranteed to identify the same parameters.

A road map to more complex dynamic models