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

“admixture”

• Example corpus

<table>
<thead>
<tr>
<th>Document 1</th>
<th>Document 2</th>
<th>Document 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>the</strong></td>
<td><strong>the</strong></td>
<td><strong>she</strong></td>
</tr>
<tr>
<td>$x_{11}$</td>
<td>$x_{21}$</td>
<td>$x_{31}$</td>
</tr>
<tr>
<td>$x_{12}$</td>
<td>$x_{22}$</td>
<td>$x_{32}$</td>
</tr>
<tr>
<td>$x_{13}$</td>
<td>$x_{23}$</td>
<td>$x_{33}$</td>
</tr>
<tr>
<td><strong>he</strong></td>
<td><strong>and</strong></td>
<td><strong>is</strong></td>
</tr>
<tr>
<td><strong>is</strong></td>
<td><strong>the</strong></td>
<td><strong>is</strong></td>
</tr>
<tr>
<td><strong>is</strong></td>
<td><strong>is</strong></td>
<td><strong>is</strong></td>
</tr>
<tr>
<td>$x_{34}$</td>
<td>$x_{34}$</td>
<td>$x_{34}$</td>
</tr>
</tbody>
</table>
Latent Dirichlet Allocation

• Generative Process

For each topic $k \in \{1, \ldots, K\}$:
$\phi_k \sim \text{Dir}(\beta)$  
[draw distribution over words]

For each document $m \in \{1, \ldots, M\}$
$\theta_m \sim \text{Dir}(\alpha)$  
[draw distribution over topics]

For each word $n \in \{1, \ldots, N_m\}$
$z_{mn} \sim \text{Mult}(1, \theta_m)$  
[draw topic assignment]
$x_{mn} \sim \phi_{z_m}$  
[draw word]

• Example corpus

<table>
<thead>
<tr>
<th>the</th>
<th>he</th>
<th>is</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
</tr>
</tbody>
</table>

Document 1

<table>
<thead>
<tr>
<th>the</th>
<th>and</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
<td>$x_{23}$</td>
</tr>
</tbody>
</table>

Document 2

<table>
<thead>
<tr>
<th>she</th>
<th>she</th>
<th>is</th>
<th>is</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{31}$</td>
<td>$x_{32}$</td>
<td>$x_{33}$</td>
<td>$x_{34}$</td>
</tr>
</tbody>
</table>

Document 3
Latent Dirichlet Allocation

- Plate Diagram
Latent Dirichlet Allocation

- Plate Diagram

\[
\begin{align*}
\text{Document-specific topic distribution} & \rightarrow \theta_m \\
\text{Topic assignment} & \rightarrow z_{mn} \\
\text{Observed word} & \rightarrow x_{mn} \\
\text{Topic} & \rightarrow \phi_k \\
\text{Dirichlet} & \rightarrow \alpha, \beta
\end{align*}
\]
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 $\phi_k$.
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 $\phi_k$
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.
**LDA for Topic Modeling**

\[ \theta_1 = \text{Dirichlet}(\theta) \]

\[ \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6 \]

- \( \phi_1 \): {Canadian gov.}
- \( \phi_2 \): {government}
- \( \phi_3 \): {hockey}
- \( \phi_4 \): {U.S. gov.}
- \( \phi_5 \): {baseball}
- \( \phi_6 \): {Japan}

(\text{Blei, Ng, & Jordan, 2003})
The 54/40’ boundary dispute is still unresolved, and Canadian and US

(Blei, Ng, & Jordan, 2003)
The 54/40' boundary dispute is still unresolved, and Canadian and US governments are involved.

\[
\phi_1 \sim \text{Dirichlet}(\alpha) \\
\phi_2 \sim \text{Dirichlet}(\beta) \\
\phi_3 \sim \text{Dirichlet}(\alpha) \\
\phi_4 \sim \text{Dirichlet}(\beta) \\
\phi_5 \sim \text{Dirichlet}(\alpha) \\
\phi_6 \sim \text{Dirichlet}(\beta)
\]

\(\theta_1\) is drawn from a Dirichlet distribution with parameters \(\alpha\) and \(\beta\).

(Blei, Ng, & Jordan, 2003)
LDA for Topic Modeling

The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard.

\[ \theta_1 = \text{Dirichlet}(\theta) \]

\[ \phi_1 \{\text{Canadian gov.}\} \]
\[ \phi_2 \{\text{government}\} \]
\[ \phi_3 \{\text{hockey}\} \]
\[ \phi_4 \{\text{U.S. gov.}\} \]
\[ \phi_5 \{\text{baseball}\} \]
\[ \phi_6 \{\text{Japan}\} \]

\[ \text{Dirichlet}(\alpha) \]

(Blei, Ng, & Jordan, 2003)
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...)

(Blei, Ng, & Jordan, 2003)
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' pitching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...
The 54/40' boundary dispute is still unresolved, and Canadian and US Coast Guard vessels regularly if infrequently detain each other's fish boats in the disputed waters off Dixon...

In the year before Lemieux came, Pittsburgh finished with 38 points. Following his arrival, the Pens finished...

The Orioles' itching staff again is having a fine exhibition season. Four shutouts, low team ERA, (Well, I haven't gotten any baseball...)

LDA for Topic Modeling

Inference and learning start with only the data

θ₁ =

θ₂ =

θ₃ =

φ₁ =

φ₂ =

φ₃ =

φ₄ =

φ₅ =

φ₆ =
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.
  2. For each topic, assign high probability to as few terms as possible.

• 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)
Case Study:
Modeling Join Imaging and Genetic data
Imaging and Genetic Data

Subject $s$

Imaging signature of Supervoxel $n$

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

Genetic loci of interest

$$G_s = \{\ell_1, \ell_3, \ell_3, \ell_5, \ell_5\}$$
Bag of Words Model

Visual Words ($I_{sn}$)

Interesting Genetic loci

Genetic Words
(Genetic variants)

$$G_s = \{l_1, l_3, l_5, l_5\}$$
Analogy: Subject as a Document

Topics (Image Patterns):

- Pattern 1
- Pattern 2
- Pattern 3

Interesting Genetic loci

Distribution of genetic variants

- Topic 1
- Topic 2
- Topic 3
Imaging – Genetic Pair Topics Signatures

\( (\mu_1, \Sigma_1) \)

\( (\mu_2, \Sigma_2) \)

\( (\mu_3, \Sigma_3) \)

\( \beta_1 \)

\( \beta_2 \)

\( \beta_3 \)

Subject Proportion

Supervoxel membership

Subject s

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 proportion

\( z_{sm}^I \sim \text{Cat}(\mathbf{1}) \)

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

\( z_{sn}^G \sim \text{Cat}(\cdot; \beta_1) \)

\( G_s = \{l_1, l_3, \overline{l_3}, l_5, l_5\} \)
Subject proportion

$z_{sm} \sim \text{Cat}(\cdot)$

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

$(\mu_k, \Sigma_k) \sim \text{NIW}(\eta^I)$

$\beta_k \sim \text{Dir}(\eta^G)$
Inference

Topic pairs

\[ p(\mu_k | \{I_{sm}\}, \{G_{sn}\}; \pi) \quad p(\beta_k | \{I_{sm}\}, \{G_{sn}\}; \pi) \]

(Imaging signature)

(Different Ranking SNPs)

\[ \pi = \{\alpha, \omega, \eta^I, \eta^G\} \] (hyper-parameters)
Factor Analysis
A road map to more complex dynamic models

- **Mixture model**
  - e.g., mixture of multinomials
  - Discrete
  - HMM
    - (for discrete sequential data, e.g., text)

- **Mixture model**
  - e.g., mixture of Gaussians
  - Continuous
  - HMM
    - (for continuous sequential data, e.g., speech signal)

- **Factor analysis**
  - Continuous

- **State space model**
  - Discrete
  - Factorial HMM
  - Switching SSM

- **Factorial HMM**
  - e.g., mixture of multinomials

- **Switching SSM**
  - e.g., mixture of Gaussians

- **HMM**
  - (for discrete sequential data, e.g., text)

- **State space model**
  - Discrete

- **Factorial HMM**
  - e.g., mixture of multinomials

- **Switching SSM**
  - e.g., mixture of Gaussians

- **HMM**
  - (for continuous sequential data, e.g., speech signal)
Recall multivariate Gaussian

- Multivariate Gaussian density:
  \[ p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]

- A joint Gaussian:
  \[ p\left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right| \mu, \Sigma) = \mathcal{N} \left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mid \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \]

- How to write down \( p(x_1), p(x_1 \mid x_2) \) or \( p(x_2 \mid x_1) \) using the block elements in \( \mu \) and \( \Sigma \)?
  - Formulas to remember:
    \[ p(x_2) = \mathcal{N} \left( x_2 \mid m_2^m, V_2^m \right) \]
    \[ m_2^m = \mu_2 \]
    \[ V_2^m = \Sigma_{22} \]
    \[ p(x_1 | x_2) = \mathcal{N} \left( x_1 \mid m_{1|2}, V_{1|2} \right) \]
    \[ m_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \]
    \[ V_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \]
Review: The matrix inverse lemma

• Consider a block-partitioned matrix:
\[ M = \begin{bmatrix} E & F \\ G & H \end{bmatrix} \]

• First we diagonalize \( M \)
\[
\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E-FH^{-1}G & 0 \\ 0 & H \end{bmatrix}
\]

• Schur complement: \( M/H = E-FH^{-1}G \)

• Then we inverse, using this formula: \( XYZ = W \Rightarrow Y^{-1} = ZW^{-1}X \)
\[
M^{-1} = \left( \begin{bmatrix} E & F \\ G & H \end{bmatrix} \right)^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \left( \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \right)^{-1} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}
\]
\[
= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}
\]

• Matrix inverse lemma
\[
(EFH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H-GE^{-1}F)^{-1}GE^{-1}
\]
Review: Some matrix algebra

• Trace and derivatives
  \[ \text{tr}[A] \overset{\text{def}}{=} \sum_i a_{ii} \]

• Cyclical permutations
  \[ \text{tr}[ABC] = \text{tr}[CAB] = \text{tr}[BCA] \]

• Derivatives
  \[ \frac{\partial}{\partial A} \text{tr}[BA] = B^T \]
  \[ \frac{\partial}{\partial A} \text{tr}[x^T Ax] = \frac{\partial}{\partial A} \text{tr}[xx^T A] = xx^T \]

• Determinants and derivatives
  \[ \frac{\partial}{\partial A} \log |A| = A^{-1} \]
Factor analysis

• An unsupervised linear regression model

\[ p(x) = \mathcal{N}(x; 0, I) \]

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

where \( \Lambda \) is called a factor loading matrix, and \( \Psi \) 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., $p(x)$) times a conditional Gaussian (e.g., $p(y|x)$) is a joint Gaussian.
- Any marginal (e.g., $p(y)$ of a joint Gaussian (e.g., $p(x,y)$) 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.)

\[
E[y] = E[\mu + \Lambda x + w] \quad \text{where} \quad w \sim N(0, \Psi)
\]

\[
= \mu + \Lambda E[x] + E[w]
\]

\[
= \mu + 0 + 0 = \mu
\]
FA = Constrained-Covariance Gaussian

• Marginal density for factor analysis (\(y\) is \(p\)-dim, \(x\) is \(k\)-dim):

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

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

\[
\text{Cov}[y] = \Lambda \Lambda^T + \Psi
\]

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

• Model
\[ p(x) = \mathcal{N}(x; 0, I) \]
\[ p(y|x) = \mathcal{N}(y; \mu + \Lambda x, \Psi) \]

• Covariance between \( x \) and \( y \)
\[
\text{Cov}[X, Y] = E[(X - 0)(Y - \mu)^T] = E[X(\mu + \Lambda X + W - \mu)^T] \\
= E[XX^T \Lambda^T + XW^T] \\
= \Lambda^T
\]

• Hence the joint distribution of \( x \) and \( y \):
\[
p\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} \mid \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda \Lambda^T + \Psi \end{bmatrix}\right)
\]

• Assume noise is uncorrelated with data or latent variables.
Inference in Factor Analysis

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

\[ \Sigma_{11} = I \]
\[ \Sigma_{12} = \Sigma_{12}^T = \Lambda^T \]
\[ \Sigma_{22} = (\Lambda \Lambda^T + \Psi) \]

we can now derive the **posterior** of the latent variable \( x \) given observation \( y \), 

\[ p(x|y) = \mathcal{N}(x|m_{1|2}, V_{1|2}) \]

, where

\[ m_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y - \mu_2) \]
\[ V_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \]
\[ = \Lambda^T (\Lambda \Lambda^T + \Psi)^{-1} (y - \mu) \]
\[ = I - \Lambda^T (\Lambda \Lambda^T + \Psi)^{-1} \Lambda \]

Applying the matrix inversion lemma

\[ (E^{-1} F(H^{-1} G)^{-1})^{-1} = E^{-1} + E^{-1} F(H^{-1} G)^{-1} E^{-1} \]

\[ (E-FH^{-1}G)^{-1} = E^{-1} + E^{-1} F(H^{-1} G)^{-1} G^{-1} \]

\[ \Rightarrow \]

\[ V_{1|2} = \left( I + \Lambda^T \Psi^{-1} \Lambda \right)^{-1} \]
\[ m_{1|2} = V_{1|2} \Lambda^T \Psi^{-1} (y - \mu) \]

• Here we only need to invert a matrix of size \(|x| \times |x|\), instead of \(|y| \times |y|\).
Geometric interpretation: inference is linear projection

- The posterior is: 
  \[ p(x|y) = \mathcal{N}(x; m_{1/2}, V_{1/2}) \]

  \[ m_{1/2} = V_{1/2} \Lambda^T \Psi^{-1} (y - \mu) \quad V_{1/2} = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} \]

- Posterior covariance does not depend on observed data \( y \)!

- 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 \( \Lambda \)
  • Manifold center \( \mu \)
  • Variance \( \Psi \)
EM for Factor Analysis

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

\[ \ell(\theta, D) = -\frac{N}{2} \log|\Lambda \Lambda^T + \Psi| - \frac{1}{2} \sum_n (y_n - \mu)^T (\Lambda \Lambda^T + \Psi)^{-1} (y_n - \mu) \]

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

• Estimating \( \mu \) is trivial: \( \hat{\mu}^{\text{ML}} = \frac{1}{N} \sum_n y_n \)

• Parameters \( \Lambda \) and \( \Psi \) are coupled nonlinearly in log-likelihood

• Complete log likelihood

\[ \ell_c(\theta, D) = \sum_n \log p(x_n, y_n) = \sum_n \log p(x_n) + \log p(y_n | x_n) \]

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

\[ = -\frac{N}{2} \log|\Psi| - \frac{1}{2} \text{tr} \left[x_n x_n^T \right] - \frac{N}{2} \text{tr}[S^{-1}], \quad \text{where } S = \frac{1}{N} \sum_n (y_n - \Lambda x_n)(y_n - \Lambda x_n)^T \]
E-step for Factor Analysis

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

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

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

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

$\langle X_n X_n^T \rangle = \text{Var}[X_n | y_n] + E[X_n | y_n] E[X_n | y_n]^T$

- Recall that we have derived:

$V_{1|2} = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1}$

$m_{1|2} = V_{1|2} \Lambda^T \Psi^{-1} (y - \mu)$

$\Rightarrow \langle X_n \rangle = m_{X_n | y_n} = V_{1|2} \Lambda^T \Psi^{-1} (y_n - \mu)$

and $\langle X_n X_n^T \rangle = V_{1|2} + m_{X_n | y_n} m_{X_n | y_n}^T$
M-step for Factor Analysis

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

\[
\frac{\partial}{\partial \Psi^{-1}} \langle \zeta \rangle = \frac{\partial}{\partial \Psi^{-1}} \left( -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \text{tr}[(X_n X_n^T)] - \frac{N}{2} \text{tr}[\langle S \rangle \Psi^{-1}] \right) = \frac{N}{2} \Psi - \frac{N}{2} \langle S \rangle \quad \Rightarrow \quad \Psi^{t+1} = \langle S \rangle
\]

\[
\frac{\partial}{\partial \Lambda} \langle \zeta \rangle = \frac{\partial}{\partial \Lambda} \left( -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \text{tr}[(X_n X_n^T)] - \frac{N}{2} \text{tr}[\langle S \rangle \Psi^{-1}] \right) = -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \Lambda} \langle S \rangle = \Psi^{-1} \sum_n \langle X_n \rangle \langle X_n^T \rangle - \Psi^{-1} \Lambda \sum_n \langle X_n X_n^T \rangle + \Lambda \sum_n \langle X_n X_n^T \rangle \Lambda^T \quad \Rightarrow \quad \Lambda^{t+1} = \left( \sum_n \langle X_n \rangle \langle X_n^T \rangle \right)^{-1} \left( \sum_n \langle X_n X_n^T \rangle \right)^{-1}
\]
Model Invariance and Identifiability

• There is *degeneracy* in the FA model.

• Since $\Lambda$ only appears as outer product $\Lambda \Lambda^T$, the model is invariant to rotation and axis flips of the latent space.

• We can replace $\Lambda$ with $\Lambda Q$ 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

Discrete $\mathcal{X}$  
Discrete $\mathcal{Y}$  
Mixture model  
e.g., mixture of multinomials
HMM  
(for discrete sequential data, e.g., text)

Continuous $\mathcal{X}$  
Continuous $\mathcal{Y}$  
Mixture model  
e.g., mixture of Gaussians
HMM  
(for continuous sequential data, e.g., speech signal)

Factorial HMM
Switching SSM

Factor analysis
State space model