Introduction to Machine Learning

Factor Analysis

Varun Chandola

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
chandola@buffalo.edu

University at Buffalo
Department of Computer Science and Engineering
School of Engeering and Applied Sciences

Moving Beyond Mixture Models

Mixture Models

- **One latent variable**

$$
\begin{aligned}
z_i &\in \{1, 2, \ldots, K\} \\
P(z_i = k) &= \pi_k \\
p(\mathbf{x}_i|\boldsymbol{\theta}) &= \sum_{k=1}^{K} p(z_i = k) p_k(\mathbf{x}_i|\boldsymbol{\theta})
\end{aligned}
$$

Moving Beyond Mixture Models

Mixture Models

- **One latent variable**

$$
\begin{aligned}
z_i &\in \{1, 2, \ldots, K\} \\
P(z_i = k) &= \pi_k \\
p(\mathbf{x}_i|\boldsymbol{\theta}) &= \sum_{k=1}^{K} p(z_i = k) p_k(\mathbf{x}_i|\boldsymbol{\theta})
\end{aligned}
$$

What if $\mathbf{z}_i \in \Re^L$?

$$
\begin{aligned}
p(\mathbf{z}_i) &= \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
p(\mathbf{x}_i|\boldsymbol{\theta}) &= \int_{\mathbf{z}_i} \boxed{p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})} p(\mathbf{z}_i) d\mathbf{z}_i
\end{aligned}
$$

Factor Analysis Models

- **Assumption**: $\mathbf{x}_i$ is a multivariate Gaussian random variable
- Mean is a function of $\mathbf{z}_i$
- Covariance matrix is fixed

$$p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- $\mathbf{W}$ is a $D \times L$ matrix (loading matrix)
- $\boldsymbol{\Psi}$ is a $D \times D$ covariance matrix
  - Assumed to be *diagonal*
- What does $\mathbf{W}$ do?

What is the Probability of $\mathbf{x}_i$

$$
\begin{aligned}
p(\mathbf{x}_i|\boldsymbol{\theta}) &= \int_{\mathbf{z}_i} p(\mathbf{x}_i|\mathbf{z}_i,\boldsymbol{\theta})p(\mathbf{z}_i)d\mathbf{z}_i \\
&= \int_{\mathbf{z}_i} \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)d\mathbf{z}_i \\
&= \mathcal{N}(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^\top)
\end{aligned}
$$

- Every $\mathbf{x}_i$ is a multivariate distribution **with same parameters**!!
- What is the mean and covariance of $\mathbf{x}$ (*dropping the subscript*)?

Simplifying Effect of Factor Analysis Model

- Often $\boldsymbol{\mu}_0$ is set to $\mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- How many parameters needed to specify the covariance?

$$
\begin{aligned}
mean(\mathbf{x}) &= \boldsymbol{\mu} \\
cov(\mathbf{x}) &= \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^\top
\end{aligned}
$$

Simplifying Effect of Factor Analysis Model

- Often $\boldsymbol{\mu}_0$ is set to $\mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}$
- How many parameters needed to specify the covariance?

$$
\begin{aligned}
mean(\mathbf{x}) &= \boldsymbol{\mu} \\
cov(\mathbf{x}) &= \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^\top
\end{aligned}
$$

- Original: D^2
- Factor analysis model: $LD + D$ (remember $\boldsymbol{\Psi}$ is a diagonal matrix)

Estimating posterior for $\mathbf{z}_i$

- What is the original intent behind LVMs?
  - Richer models of $p(\mathbf{x})$
- But they can also be used as a lower dimensional representation of $\mathbf{x}$.
- Mixture models?
- Factor analysis model?
  - What is $p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta})$?

Estimating posterior for $\mathbf{z}_i$

- What is the original intent behind LVMs?
  - Richer models of $p(\mathbf{x})$
- But they can also be used as a lower dimensional representation of $\mathbf{x}$.
- Mixture models?
- Factor analysis model?
  - What is $p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta})$?

$$
\begin{aligned}
p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{m}_i, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &\triangleq (\boldsymbol{\Sigma}_0^{-1} + \mathbf{W}^\top \boldsymbol{\Psi}^{-1}\mathbf{W})^{-1} \\
\mathbf{m}_i &\triangleq \boldsymbol{\Sigma}(\mathbf{W}^\top \boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)
\end{aligned}
$$

Interpreting Latent Factors

- Each $\mathbf{x}_i$ has a corresponding $\mathbf{z}_i$
- Each $\mathbf{z}_i$ is a multivariate Gaussian random variable with mean $\mathbf{m}_i$ (A $L \times 1$ vector)
- One can "embed" $\mathbf{x}_i$ ($D \times 1$ vector) into a $L \times 1$ space

Issue of Unidentifiability

- Consider an orthogonal rotation matrix $\mathbf{R}$

$$\mathbf{R}\mathbf{R}^\top = \mathbf{I}$$

- Let $\widehat{\mathbf{W}} = \mathbf{W}\mathbf{R}$
- The FA model with $\widehat{\mathbf{W}}$ will also have the same result, i.e., the pdf of observed $\mathbf{x}$ will still be the same
- Thus FA model can have multiple solutions
- The predictive power of the model does not change
- But intepreting latent factors can be an issue

Learning Parameters

- FA model parameters: $\mathbf{W}, \boldsymbol{\Psi}, \boldsymbol{\mu}$
  - $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ can be "absorbed" in $\mathbf{W}$ and μ, respectively
- A simple extension of the mixture model EM algorithm will work here

Factor Analysis - A Real World Example

- 2004 Cars Data
- Original - 11 features
- Factor analysis results in 2 factors

Variants of Factor Analysis

- If we use a non-gaussian distribution for $p(\mathbf{z}_i)$ we arrive at *Independent Component Analysis*.
- If $\mathbf{\Psi} = \sigma^2\mathbf{I}$ and $\mathbf{W}$ is orthonormal $\Rightarrow$ FA is equivalent to **Probabilistic Principal Components Analysis** (PPCA)
- If $\sigma^2 \to 0$, FA is equivalent to PCA
- What is PCA?

References