Multivariate Statistics

Lecture 13

Fudan University

Outline

Outline

Factor Analysis

Let the observable vector $\mathbf{t}$ be written as

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\mathbf{t}$, $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$ are column vectors of d components, $\mathbf{x}$ is column vector of q components $(q \leq d)$, and $\mathbf{W}$ is a $d \times q$ matrix.

We assume $\boldsymbol{\epsilon}$ is distributed independently of $\mathbf{x}$ and with mean $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and covariance matrix $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\Psi}$ is diagonal.

1. The model is similar to regression, but $\mathbf{x}$ is unobserved.
2. There are two kinds of models:
   - $\mathbf{x}$ is a nonrandom vector
   - $\mathbf{x}$ is a random vector: $\mathbf{t}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha$

Factor Analysis

Example of mental tests for $\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$:

1. Each component of $\mathbf{t}$ is a (centralized) score on a battery of tests.
2. The components of $\mathbf{x}$ are the scores of the mental factors, linear combinations of these enter into the test scores.
3. Each component of $\boldsymbol{\mu}$ is the average score in the population.
4. The coefficients of these linear combinations are the elements of $\mathbf{W}$, and these are called factor loadings (common factors).
5. A component of $\boldsymbol{\epsilon}$ is the part of the test score not "explained" by the common factors (error).

Factor Analysis

Example of recommending system for $\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$:

1. Each component of $\mathbf{t}$ is a (centralized) score on an item.
2. The components of $\mathbf{x}$ are the attributes of the user.
3. Each component of $\boldsymbol{\mu}$ is the average score in the population.
4. The coefficients of these linear combinations are the elements of $\mathbf{W}$, and these are called factor loadings (common factors).
5. The components of $\boldsymbol{\epsilon}$ are noise.

Factor Analysis

The columns of $\mathbf{W} \in \mathbb{R}^{d \times q}$ establish an q-dimensional subspace of $\mathbb{R}^d$.

1. This subspace is called the factor space.
2. Vector $\mathbf{x} \in \mathbb{R}^q$ can be viewed as coordinates of a point in factor space.

Factor Analysis

There is a indeterminacy in the model.

1. Suppose $\mathbf{\Phi} = \mathbf{I}$, $\mathbf{x}^* = \mathbf{C}^{-1}\mathbf{x}$, $\mathbf{W}^* = \mathbf{W}\mathbf{C}$, where $\mathbf{C} \in \mathbb{R}^{q \times q}$ is orthogonal, then $\mathbf{t} = \mathbf{W}^*\mathbf{x}^* + \boldsymbol{\mu} + \boldsymbol{\epsilon}$ and $\mathbb{E}\left[\mathbf{x}^*\mathbf{x}^{*\top}\right] = \mathbf{I}$.
2. To identify the parameters, we require additional assumption such as $\mathbf{\Gamma} = \mathbf{\Lambda}^\top \mathbf{\Psi}^{-1} \mathbf{\Lambda}$ is diagonal.

Outline

Probabilistic Principle Component Analysis

Let $\mathbf{t}_1, \ldots, \mathbf{t}_N$ be N independent observation and we have

$$\mathbf{t}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \boldsymbol{\epsilon}_\alpha,$$

where $\mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon}_\alpha \sim \mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I})$ are independent.

Then, we have $\mathbf{t}_\alpha \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

The log-likelihood function is

$$-\frac{Nd\ln(2\pi)}{2} - N\ln\det(\mathbf{C}) - \mathrm{tr}\Big(\mathbf{C}^{-1}\sum_{\alpha=1}^{N}(\mathbf{t}_\alpha - \boldsymbol{\mu})(\mathbf{t}_\alpha - \boldsymbol{\mu})^\top\Big).$$

The Maximum Likelihood Estimators

The maximum likelihood estimators of $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 are

$$\boldsymbol{\mu} = \bar{\mathbf{t}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{t}_\alpha, \quad \hat{\mathbf{W}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \hat{\sigma}^2\mathbf{I})\mathbf{R} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j,$$

where $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ with columns are the principal eigenvectors of

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{t}_\alpha - \bar{\mathbf{t}})(\mathbf{t}_\alpha - \bar{\mathbf{t}})^\top,$$

$\boldsymbol{\Lambda}_q \in \mathbb{R}^{q\times q}$ is diagonal matrix with corresponding eigenvalues $\lambda_1, \dots, \lambda_q$ and $\mathbf{R}$ is any $q \times q$ orthogonal matrix.

The Maximum Likelihood Estimators

The MLE estimator also minimize the Frobenius norm error

$$(\hat{\mathbf{W}}, \hat{\sigma}^2) = \operatorname*{arg\,min}_{\mathbf{W}\in\mathbb{R}^{d\times q}, \sigma^2\in\mathbb{R}^+} \left\| \hat{\mathbf{\Sigma}} - \left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right) \right\|_F.$$

Lemma 1

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ and $q = \min\{m, n\}$. Define the diagonal matrix $\mathbf{\Sigma}(\mathbf{A})$ whose (i, i)-th element is the i-th singular value of $\mathbf{A}$ and the others are zero. We define $\mathbf{\Sigma}(\mathbf{A})$. Then we have

$$\|\mathbf{A} - \mathbf{B}\| \geq \|\mathbf{\Sigma}(\mathbf{A}) - \mathbf{\Sigma}(\mathbf{B})\|.$$

for every unitarily invariant norm.

The EM Algorithm

For the model

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\mathbf{x} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I})$ are independent.

View $\{\mathbf{x}_\alpha\}_{\alpha=1}^N$ as missing data and $\{\mathbf{x}_\alpha, \mathbf{t}_\alpha\}_{\alpha=1}^N$ as the complete data.

1. $\mathbf{t} \,|\, \mathbf{x} \sim \mathcal{N}_d(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
2. $\mathbf{x} \,|\, \mathbf{t} \sim \mathcal{N}_q(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$, where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$

The EM Algorithm

The update of the EM algorithm

1. In E-step, we take the expectation

$$l_C = \mathbb{E}\left[\ln\left(\prod_{\alpha=1}^{N} p(\mathbf{x}_\alpha \,|\, \mathbf{t}_\alpha)\right)\right].$$

2. In the M-step, we maximized l_C with respect to $\mathbf{W}$ and σ^2:

$$\begin{aligned}\tilde{\mathbf{W}} =& \hat{\mathbf{\Sigma}}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\mathbf{\Sigma}}\mathbf{W})^{-1},\\ \tilde{\sigma}^2 =& \frac{1}{d}\mathrm{tr}\left(\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^\top\right).\end{aligned}$$

Note that the computational complexity of EM is $\mathcal{O}(Ndq)$, while MLE requires $\mathcal{O}(Nd^2 + d^3)$.