EECS 275 Matrix Computation

Ming-Hsuan Yang

Electrical Engineering and Computer Science
University of California at Merced
Merced, CA 95344
http://faculty.ucmerced.edu/mhyang

Lecture 8

Overview

- Multivariate Gaussian
- Mahalanobis distance
- Probabilistic PCA
- Factor analysis

Reading

- Chapter 7 and 9 of *Principal Component Analysis* by Ian Jolliffe

Multivariate Gaussian distribution

- The d-dimensional Gaussian distribution of $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}, \mathcal{C}) &= \frac{1}{(2\pi)^{d/2}|\mathcal{C}|^{1/2}} \exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathcal{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \\ &= \frac{1}{(2\pi)^{d/2}|\mathcal{C}|^{1/2}} \exp(-\tfrac{1}{2}\Delta^2) \end{aligned}$$

  where $\boldsymbol{\mu}$ is the mean and $\mathcal{C}$ is the covariance matrix

- Assume independent observations, find $\boldsymbol{\mu}$ and $\mathcal{C}$ that maximize log likelihood from a set of n points, $\mathbf{x}_1, \ldots, \mathbf{x}_n$

$$\begin{aligned} p(X|\boldsymbol{\mu}, \mathcal{C}) &= \textstyle\prod_{i=1}^n p(\mathbf{x}_i|\boldsymbol{\mu}, \mathcal{C}) \\ \mathcal{L} &= \log \textstyle\prod_{i=1}^n p(\mathbf{x}_i|\boldsymbol{\mu}, \mathcal{C}) \\ &= -\tfrac{nd}{2}\log(2\pi) - \tfrac{n}{2}\log|\mathcal{C}| - \tfrac{1}{2}\textstyle\sum_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \mathcal{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \end{aligned}$$

- Maximum likelihood estimate:

$$\begin{aligned} \tfrac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = 0 \quad &\Rightarrow \quad \hat{\boldsymbol{\mu}} = \tfrac{1}{n}\textstyle\sum_i \mathbf{x}_i \quad \text{(sample mean)} \\ \tfrac{\partial \mathcal{L}}{\partial \mathcal{C}} = 0 \quad &\Rightarrow \quad \hat{\mathcal{C}} = \tfrac{1}{n}\textstyle\sum_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \quad \text{(sample covariance)} \end{aligned}$$

Properties of Gaussian distribution

$$p(\mathbf{x}|\boldsymbol{\mu}, \mathcal{C}) = \frac{1}{(2\pi)^{d/2}|\mathcal{C}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathcal{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$$

- The ellipsoid that best represents the distribution of data points can be estimated by the covariance matrix $\mathcal{C}$
- Marginal densities (obtained by integrating out some of the variables) are themselves Gaussian
- Conditional densities (by setting some variables to fixed values) are also Gaussian
- Can find a linear transformation which diagonalizes $\mathcal{C}$ so that the density function can be factorized

$$\mathcal{C} = \sigma^2 I, \quad p(\mathbf{x}|\boldsymbol{\mu}, \mathcal{C}) = \prod_{i=1}^{n} p(x_i|\mu_i, \sigma_i)$$

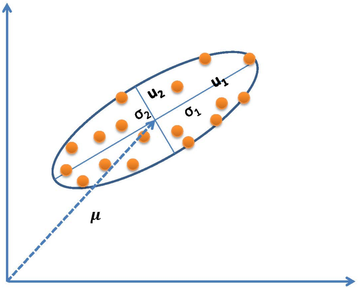
- For given values of $\boldsymbol{\mu}$ and $\mathcal{C}$, the Gaussian density function maximizes the entropy
- Useful for linear classifiers (e.g., Fisher linear discriminant)

Geometric interpretation

- The equi-density contours of a non-singular Gaussian (i.e., $P(\mathbf{x}|\boldsymbol{\mu}, \mathcal{C}) = k$) where k is a constant) are ellipsoids (i.e., linear transformation of hyperspheres)
- The directions of the principal axes of the ellipsoids are the eigenvectors $\mathbf{u}$ of covariance matrix $\mathcal{C}$, and the lengths are the corresponding singular values σ ($\sigma_i = \sqrt{\lambda_i}$ where λ_i is an eigenvalue)

$$\mathcal{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

- For 2D case,

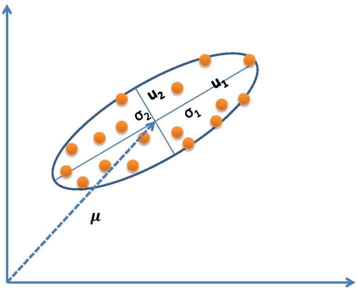


Geometric interpretation

- Let $\mathcal{C} = U\Sigma U^\top = (U\Sigma^{1/2})(U\Sigma^{1/2})^\top$ (i.e., eigendecomposition) where the columns of U are orthonormal basis and Σ is a diagonal matrix

$$X \sim N(\boldsymbol{\mu}, \mathcal{C}) \iff X \sim \boldsymbol{\mu} + U\Sigma^{1/2}N(0, I) \iff X \sim \boldsymbol{\mu} + UN(0, \Sigma)$$

- The distribution of $N(\boldsymbol{\mu}, \mathcal{C})$ is equivalent to $N(0, I)$ scaled by $\Sigma^{1/2}$, rotated by U and translated by $\boldsymbol{\mu}$
- For 2D case,



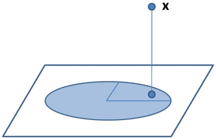
Mahalanobis distance

- The quantity

$$d_M^2 = \Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (C^{-1/2}(\mathbf{x} - \boldsymbol{\mu}))^\top (C^{-1/2}(\mathbf{x} - \boldsymbol{\mu}))$$

  is called the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$
- Also known as generalized squared inter-point distance
- The distance of a point $\mathbf{x}$ to the center of mass divided by the width of the ellipsoid in the direction of $\mathbf{x}$
- Linear transformation of the coordinate system
- Keep its quadratic form and remain non-negative
- If $\mathcal{C} = I$, Mahalanobis distance reduces to Euclidean distance
- If $\mathcal{C}$ is diagonal, the resulting distance is normalized Euclidean distance $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^m \frac{(\mathbf{x}_i - \mathbf{y}_i)^2}{\sigma_i^2}}$ where σ_i is the standard deviation of $\mathbf{x}_i$
- Can be approximated with eigenvectors of $\mathcal{C}$
- Related to similarity learning or metric learning

Generative PCA model



- A subspace is spanned by the orthonormal basis (eigenvectors computed from covariance matrix)
- Can interpret each observation with a generative model
- Estimate (approximately) the probability of generating each observation with Gaussian distribution, $p(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$
- Several ways to approximate $p(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, e.g., distance to subspace, distance within subspace, and combination
- Each observation has a projected latent variable
- Used in object modeling, detection, tracking, recognition, etc.

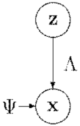
Factor analysis

- A generative dimensionality reduction algorithm
- Let $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{x}$ is modeled by $\mathbf{z}$, dubbed as factors ($d < m$)

$$\mathbf{x} = \Lambda \mathbf{z} + \varepsilon$$

  - Λ is factor loading matrix
  - $\mathbf{z}$ is assumed be $N(0, I)$ distributed (zero mean, unit variance normals)
  - The factors $\mathbf{z}$ model correlation between the elements of $\mathbf{x}$
  - ε is a random variable to account for noise and assumed to be distributed with $N(0, \Psi)$ where Ψ is a diagonal matrix (whereas PCA uses an isotropic error model with $\psi_i = \sigma^2$)
  - ε accounts for independent noise in each element of $\mathbf{x}$
  - The diagonality of Ψ is a key assumption: constraining the error covariance Ψ for estimation
  - The observed variable, $\mathbf{x}_i$, are conditionally independent given the factors $\mathbf{z}$
  - $\mathbf{x}$ is $N(0, \Lambda\Lambda^\top + \Psi)$ distributed (whereas PCA models with $N(0, \Lambda\Lambda^\top + \sigma^2 I)$

Properties of factor analysis



- Factor analysis: $\mathbf{x} = \Lambda z + \varepsilon$
- Latent variables $\mathbf{z}$: explain correlations between $\mathbf{x}$
- ε_i represents variability unique to a particular $\mathbf{x}_i$
- Differ from PCA which treats covariance and variance identically
- Want to infer Λ and Ψ from $\mathbf{x}$
- Suppose Λ and Ψ are known, by linear projection

$$E[\mathbf{z}|\mathbf{x}] = \beta \mathbf{x}$$

where $\beta = \Lambda^\top (\Psi + \Lambda\Lambda^\top)^{-1}$, since the joint Gaussian of data $\mathbf{x}$ and factors $\mathbf{z}$:

$$p(\begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}) = N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda\Lambda^\top + \Psi & \Lambda \\ \Lambda^\top & I \end{bmatrix})$$

Properties of factor analysis (cont'd)

- Note that since Ψ is diagonal, using matrix inversion lemma

$$(\Psi + \Lambda\Lambda^\top)^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda(I + \Lambda^\top\Psi^{-1}\Lambda)^{-1}\Lambda^\top\Psi^{-1}$$

- The second moment of factors:

$$\begin{aligned} E[\mathbf{z}\mathbf{z}^\top|\mathbf{x}] &= Var(\mathbf{z}|\mathbf{x}) + E[\mathbf{z}|\mathbf{x}]E[\mathbf{z}|\mathbf{x}]^\top \\ &= I - \beta\Lambda + \beta\mathbf{x}\mathbf{x}^\top\beta^\top \end{aligned}$$

  where $\beta = \Lambda^\top(\Psi + \Lambda\Lambda^\top)^{-1}$

- Expectation of first and second moments provide measure of uncertainty in the factors, which PCA does not have
- Ψ and Λ can be computed by the EM algorithm

EM algorithm for factor analysis

- Expectation-Maximization: technique for dealing with missing data
- Start with some initial guess of missing data and evaluate the expected values
- Optimize the missing parameters by taking derivative of likelihood of observed and missing data w.r.t. parameters
- Repeat until the data likelihood does not change
- E-step: Given Λ and Ψ, for each data point $\mathbf{x}_i$, compute

$$\begin{aligned} E[\mathbf{z}|\mathbf{x}] &= \beta\mathbf{x} \\ E[\mathbf{z}\mathbf{z}^\top|\mathbf{x}] &= Var(\mathbf{z}|\mathbf{x}) + E[\mathbf{z}|\mathbf{x}]E[\mathbf{z}|\mathbf{x}]^\top \\ &= I - \beta\Lambda + \beta\mathbf{x}\mathbf{x}^\top\beta^\top \end{aligned}$$

- M-step:

$$\begin{aligned} \Lambda^{new} &= \left(\textstyle\sum_{i=1}^n \mathbf{x}_i E[\mathbf{z}|\mathbf{x}_i]^\top\right)\left(\textstyle\sum_{i=1}^n E[\mathbf{z}\mathbf{z}^\top|\mathbf{x}_i]\right)^{-1} \\ \Psi^{new} &= \frac{1}{n}\text{diag}\{\textstyle\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top - \Lambda^{new}E[\mathbf{z}|\mathbf{x}_i]\mathbf{x}_i^\top\} \end{aligned}$$

where diag operator sets all off-diagonal elements to zero

FA and PCA

- Factor analysis provides a proper probabilistic model
- PCA is rotationally invariant; FA is not
- Given a set of data points, would Λ correspond to orthonormal basis of a PCA subspace?
- No, in most cases
- However, Λ corresponds to orthonormal basis if FA has isotropic error model, i.e., $\psi_i = \sigma^2$

Probabilistic principal component analysis

- Let $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^d$, from factor analysis we have $\mathbf{x} = \Lambda\mathbf{z} + \varepsilon$, with isotropic noise model $N(0, \sigma^2 I)$
- The conditional probability of $\mathbf{x}$ given $\mathbf{z}$ is given by

$$\mathbf{x}|\mathbf{z} \sim N(\Lambda z, \sigma^2 I)$$

- Since $\mathbf{z} \sim N(0, I)$, marginal distribution for $\mathbf{x}$ is

$$\mathbf{x} \sim N(0, \widetilde{C})$$

  where $\widetilde{C} = \Lambda\Lambda^\top + \sigma^2 I$
- Log likelihood of data

$$\mathcal{L} = -\frac{n}{2}\{m \ln(2\pi) + \ln|\widetilde{C}| + \mathrm{tr}(\widetilde{C}^{-1} S)\}$$

  where

$$S = \frac{1}{n} X X^\top$$

- Estimating Λ and σ^2 can be obtained by maximizing $\mathcal{L}$ using the EM algorithm similar to that in factor analysis

Probabilistic principal component analysis (cont'd)

- Maximize log likelihood with the EM algorithm,

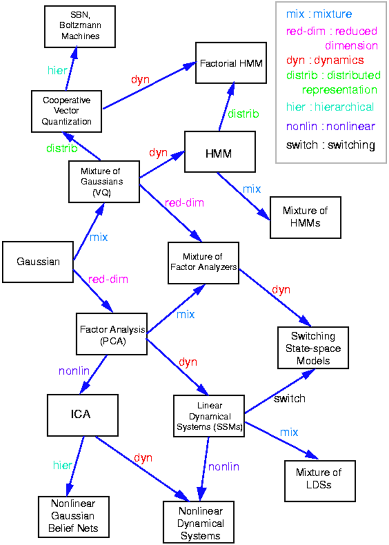
$$\Lambda = U(\Sigma - \sigma^2 I)^{1/2} R$$

  - $U_{m \times d}$ is the first d eigenvectors computed from covariance matrix S
  - $\Sigma_{d \times d}$ is a diagonal matrix corresponding to the first d eigenvalues, λ_i
  - $R_{d \times d}$ is an arbitrary orthogonal rotation matrix (note $\mathbf{z}$ has a uniform Gaussian distribution)
  - The noise variance σ^2 is the residual variance per dimension

$$\sigma^2 = \frac{1}{m-d} \sum_{i=d+1}^{m} \lambda_i$$

  - See "Probabilistic Principal Component Analysis," by Tipping and Bishop for details

Big picture



"A unifying review of linear Gaussian models" [Ghahramani and Roweis]