Machine Learning Homework 2*

Hongzhou Liu
517030910214
deanlhz@sjtu.edu.cn

1 PCA algorithm

2 Factor Analysis (FA)

By Bayesian formula, we know that

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} \tag{1}$$

Here,

$$p(\mathbf{x}) = p(\mathbf{A}\mathbf{y} + \mu + \mathbf{e}) \tag{2}$$

and

$$p(\mathbf{x}|\mathbf{y}) = G(\mathbf{x}|\mathbf{A}\mathbf{y} + \mu, \Sigma_e), p(\mathbf{y}) = G(\mathbf{y}|0, \Sigma_y) \tag{3}$$

generally

$$p(\mathbf{e}) = G(\mathbf{e}|\mu_e, \Sigma_e) \tag{4}$$

Here, $\mathbf{A}\mathbf{y} + \mu$ is an affine transformation of $\mathbf{y}$, thus

$$p(\mathbf{x}) = G(\mathbf{A}\mathbf{y} + \mu|\mu, \mathbf{A}\Sigma_y\mathbf{A}^T) + G(\mathbf{e}|\mu_e, \Sigma_e) = G(\mathbf{x}|\mu + \mu_e, \mathbf{A}\Sigma_y\mathbf{A}^T + \Sigma_e) \tag{5}$$

Then,

$$p(\mathbf{y}|\mathbf{x}) = \frac{G(\mathbf{x}|\mathbf{A}\mathbf{y} + \mu, \Sigma_e)G(\mathbf{y}|0, \Sigma_y)}{G(\mathbf{x}|\mu + \mu_e, \mathbf{A}\Sigma_y\mathbf{A}^T + \Sigma_e)} \tag{6}$$

The density function of Gaussian distribution is

$$G(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}}\exp(-\frac{1}{2}(\mathbf{x} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)) \tag{7}$$

k is the dimension of $\mathbf{x}$. Then we consider the exponential terms of $p(\mathbf{y}|\mathbf{x})$ which is

$$-\frac{1}{2}(\mathbf{x}-\mathbf{A}\mathbf{y}-\mu)^T\Sigma_e^{-1}(\mathbf{x}-\mathbf{A}\mathbf{y}-\mu)-\frac{1}{2}\mathbf{y}^T\Sigma_y^{-1}\mathbf{y}+\frac{1}{2}(\mathbf{x}-\mu+\mu_e)^T(\mathbf{A}\Sigma_y\mathbf{A}^T+\Sigma_e)^{-1}(\mathbf{x}-\mu+\mu_e) \tag{8}$$

We only consider terms containing $\mathbf{y}$, that is

$$\begin{aligned} &-\frac{1}{2}[-\mathbf{x}^T\Sigma_e^{-1}\mathbf{A}\mathbf{y} - \mathbf{y}^T\mathbf{A}^T\Sigma_e^{-1}(\mathbf{x} - \mathbf{A}\mathbf{y} - \mu) + \mu^T\Sigma_e^{-1}\mathbf{A}\mathbf{y} + \mathbf{y}^T\Sigma_y^{-1}\mathbf{y}] \\ &= -\frac{1}{2}[(\mu - \mathbf{x})^T\Sigma_e^{-1}\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}^T\Sigma_e^{-1}(\mu - \mathbf{x}) + \mathbf{y}^T(\mathbf{A}^T\Sigma_e^{-1}\mathbf{A} + \Sigma_y^{-1})\mathbf{y}] \end{aligned} \tag{9}$$

We know that

$$\begin{aligned} &(\mathbf{y} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mu) \\ &= \mathbf{y}^T\Sigma^{-1}\mathbf{y} - \mathbf{y}^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}\mathbf{y} + \mu^T\Sigma^{-1}\mu \end{aligned} \tag{10}$$

*GitHub repo: https://github.com/DeanAlkene/CS420-MachineLearning/tree/master/A2

Compare 9 and 10 we get,

$$\Sigma_{\mathbf{y}|\mathbf{x}} = (\mathbf{A}^T\Sigma_e^{-1}\mathbf{A} + \Sigma_y^{-1})^{-1} \tag{11}$$

and

$$\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}\mu_{\mathbf{y}|\mathbf{x}} = \mathbf{A}^T\Sigma_e^{-1}(\mathbf{x} - \mu) \tag{12}$$

Hence

$$p(\mathbf{y}|\mathbf{x}) = G(\mathbf{y}|(\mathbf{A}^T\Sigma_e^{-1}\mathbf{A} + \Sigma_y^{-1})^{-1}\mathbf{A}^T\Sigma_e^{-1}(\mathbf{x} - \mu), (\mathbf{A}^T\Sigma_e^{-1}\mathbf{A} + \Sigma_y^{-1})^{-1}) \tag{13}$$

3 Independent Component Analysis (ICA)

In ICA, we have a linear combination of source vectors $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{s}$ are independent sources. The goal is to find a transformation $\mathbf{W}$ to seperate each sources into $\mathbf{y}$ and make each entry in $\mathbf{y}$ as independent as possible.

The Central Limit Theorem tells us that a sum of independent random variables from arbitrary distributions tends towards a Gaussian distribution, under certain conditions. Let's consider ICA as

$$\mathbf{y} = \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{A}\mathbf{s} = (\mathbf{w}^T\mathbf{A})\mathbf{s} = \mathbf{z}^T\mathbf{s} \tag{14}$$

Now, $\mathbf{y}$ is a liner combination of random variables $\mathbf{s}$. According to the Central Limit Theorem, $\mathbf{y}$ should be closer to Gaussian than any s_i in $\mathbf{s}$. However, to pursue independence among each entry of $\mathbf{y}$, we ought to minimize affects of being closer to Gaussian brought by $\mathbf{z}^T$. It is equally to say, we should take $\mathbf{w}$ that maximizes the non-Gaussianity, which is a principle for ICA estimation.

In another perspective, let's prove that in ICA at most one Gaussian variable is allowed. Let's consider $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{s} = s_1, s_2$. Without lossing of generality, let $\mathbf{s} \sim \mathcal{N}(0, I)$. Then,

$$\mathbf{x} \sim \mathcal{N}(0, \mathbf{A}\mathbf{A}^T) \tag{15}$$

Here is an orthogonal transformation matrix $\mathbf{R}$. Apply it on $\mathbf{A}$ as $\mathbf{A}' = \mathbf{A}\mathbf{R}$, we have

$$\mathbf{x}' = \mathbf{A}\mathbf{R}\mathbf{s} \sim \mathcal{N}(0, \mathbf{A}\mathbf{R}\mathbf{R}^T\mathbf{A}^T) = \mathcal{N}(0, \mathbf{A}\mathbf{A}^T) \tag{16}$$

Thus, due to the symetric property of multivariable Gaussian, we cannot tell the source $\mathbf{s}$ from the observation $\mathbf{x}$ because there're infinite much $\mathbf{s}$. In this way, we also proved that, to implement ICA we should stay away from Gaussian.

4 Dimension Reduction by FA

5 Spectral clustering

References

[1] J. Yang and L. Jin, "An Improved RPCL Algorithm for Determining Clustering Number Automatically," TENCON 2006 - 2006 IEEE Region 10 Conference, Hong Kong, 2006, pp. 1-3.