A Proportional Structure of the One Dimensional Solutions to Linear Dimensionality Reduction with the Objective of Maximizing *f*-Divergences

Sihui Wang

1. Introduction

We have found out that there is a proportional structure of the one dimensional solutions to linear dimensionality reduction with the objective of maximizing Fisher information distance. It turns out that our methods could also be applied to the cases of KL-divergence, symmetric KL divergence, and Hellinger distance. Basically the solutions in these three cases also have the same proportional structure.

2. KL Divergence

Notations

Suppose $X \sim \mathcal{N}(0, I_n)$, $Y \sim \mathcal{N}(\mu, \Sigma)$, $A \in \mathbb{R}^{1\times n}$, $AA^T = 1$, then $AX \sim \mathcal{N}(0,1)$, $AY \sim \mathcal{N}(A\mu, A\Sigma A^T)$.

Assume that $\Sigma = V\Lambda V^T$, where $\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ is a diagonal matrix with $\lambda_1, \dots, \lambda_n$ being the eigenvalues of Σ, and $V = (v_1 \quad \cdots \quad v_n)$ is an orthogonal matrix with $v_1, \dots, v_n \in \mathbb{R}^{n\times 1}$ being the eigenvectors of Σ.

Assume that $\mu = V\alpha = (v_1 \quad \cdots \quad v_n)\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \sum_i \alpha_i v_i$ with $\alpha_1, \dots, \alpha_n$ being scalars.

Assume that $A = \beta^T V^{\mathrm{T}} = (\beta_1 \quad \cdots \quad \beta_n)\begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} = \sum_i \beta_i v_i^T$ with $\beta_1, \dots, \beta_n$ being scalars.

Since V is orthogonal, the constraint $AA^T = 1$ is equivalent to the constraint of $\beta^{\mathrm{T}}\beta = 1$.

Now we have:

$$A\Sigma A^T = \beta^T \Lambda \beta$$
$$\mu^T A^T A \mu = \alpha^T \beta \beta^T \alpha$$
$$\alpha^T \alpha = \mu^T \mu$$

Proportional Structure of Solutions

We have:

$$D_{KL}(\mathcal{N}(0,1) \| \mathcal{N}(A\mu, A\Sigma A^T)) = \frac{1}{2}\log(A\Sigma A^T) + \frac{1 + \mu^T A^T A \mu}{2A\Sigma A^T} - \frac{1}{2}$$

In order to find $A = \arg\max_{AA^T=1} D_{KL}(AX \| AY)$, it is equivalent for us to find $\beta \in \mathbb{R}^{n\times 1}$ so that:

$$\beta = \arg\max_{\beta^T\beta=1} \left[\log(\beta^T \Lambda \beta) + \frac{1 + \alpha^T \beta \beta^T \alpha}{\beta^T \Lambda \beta}\right] \quad ①$$

Note that $\alpha^T \beta \beta^T \alpha = \|\mu\|_2^2 \cos^2\theta$ where θ is the angle between the vector of α and the vector of β.

Now consider a family of constrained optimization problems:

$$\beta(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\max} \left[\log(\beta^T\Lambda\beta) + \frac{1+\|\mu\|_2^2\cos^2\theta}{\beta^T\Lambda\beta}\right] \quad ②$$

If β^* is the optimal solution to ①, then it is necessary that for some θ, β^* is the optimal solution to ②.

Note that under the constraint of ②, $1+\|\mu\|_2^2\cos^2\theta$ is a constant positive number. For any constant positive number C, the function $f(x) = \log x + \frac{C}{x}(x>0)$ is monotonously decreasing in the interval of $(0,C]$ and monotonously increasing in the interval of $[C,\infty)$, which means that $f(x)$ will reach its maximum only when x reaches either its minimum or its maximum under the constraints. This means that in order to obtain the optimal solution to ②, it is necessary that $\beta^T\Lambda\beta$ reach either the maximum or minimum under the constraints of ②.

So, if for some θ, β^* is the optimal solution to ②, it is necessary that for the same θ, β^* is the optimal solution to either of the following two problems:

$$\beta^+(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\max} \beta^T\Lambda\beta \quad ③$$

$$\beta^-(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\min} \beta^T\Lambda\beta \quad ④$$

So, for Lagrange function:

$$L(\beta, l_1, l_2) = \beta^{\mathrm{T}}\Lambda\beta + l_1(\beta^T\beta - 1) + l_2(\tilde{\alpha}^T\beta - cos\theta)$$

where $\tilde{\alpha} = \frac{\alpha}{\|\alpha\|_2}$, it is necessary that $\frac{\partial L(\beta,l_1,l_2)}{\partial\beta} = 2\Lambda\beta + 2l_1\beta + l_2\tilde{\alpha} = 0$.

So we have:

$$(\Lambda + l_1 I_n)\beta = -\frac{l_2}{2\|\alpha\|_2}\alpha$$

which means that for some parameters γ, k, we have:

$$(\lambda_i + \gamma)\beta_i = k\alpha_i (1 \le i \le n)$$

Hence, we find out that in $\beta = (\beta_1, \dots, \beta_n)^{\mathrm{T}}$, $\beta_1, \dots, \beta_n$ are proportional to each other:

$$\beta_1 : \cdots : \beta_n = \frac{\alpha_1}{\lambda_1+\gamma} : \cdots : \frac{\alpha_n}{\lambda_n+\gamma}$$

Since $\beta^T\beta = 1$, β will be uniquely determined if we could decide γ.

3. Symmetric KL Divergence

In order to find $A = \underset{AA^T=1}{\arg\max} D_{SKL}(AX\|AY)$, it is equivalent for us to find $\beta \in \mathbb{R}^{n\times 1}$ so that:

$$\beta = \underset{\beta^T\beta=1}{\arg\max}\left[\alpha^T\beta\beta^T\alpha + \beta^T\Lambda\beta + \frac{1+\alpha^T\beta\beta^T\alpha}{\beta^T\Lambda\beta}\right] \quad ⑤$$

If β^* is the optimal solution to ⑤, then it is necessary that for some θ, β^* is the optimal solution to the following:

$$\beta(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\max} \left[\|\mu\|_2^2\cos^2\theta + \beta^T\Lambda\beta + \frac{1+\|\mu\|_2^2\cos^2\theta}{\beta^T\Lambda\beta}\right] \quad ⑥$$

Note that $\|\mu\|_2^2 \cos^2\theta$ is a constant positive number under the constraint of ⑥. For any constant positive number C, the function $f(x) = x + \frac{C}{x}(x > 0)$ is monotonously decreasing in the interval of $(0, \sqrt{C}]$ and monotonously increasing in the interval of $[\sqrt{C}, \infty)$, which means that if for some θ, β^* is the optimal solution to ⑥, it is necessary that for the same θ, β^* is the optimal solution to either of the following two problems:

$$\beta^+(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\max} \beta^T\Lambda\beta \quad ⑦$$

$$\beta^-(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\min} \beta^T\Lambda\beta \quad ⑧$$

Following Lagrange multiplier method as before, we could obtain that the solutions have the same proportional structure.

4. Hellinger Distance

In order to find $A = \underset{AA^T=1}{\arg\max} D_H(AX\|AY)$, it is equivalent for us to find $\beta \in \mathbb{R}^{n\times 1}$ so that:

$$\beta = \underset{\beta^T\beta=1}{\arg\max}\left[2\log(\sqrt{\beta^T\Lambda\beta} + \frac{1}{\sqrt{\beta^T\Lambda\beta}}) + \frac{\alpha^T\beta\beta^T\alpha}{1+\beta^T\Lambda\beta}\right] \quad ⑨$$

If β^* is the optimal solution to ⑨, then it is necessary that for some θ, β^* is the optimal solution to the following:

$$\beta(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\max}\left[2\log(\sqrt{\beta^T\Lambda\beta} + \frac{1}{\sqrt{\beta^T\Lambda\beta}}) + \frac{\|\mu\|_2^2\cos^2\theta}{1+\beta^T\Lambda\beta}\right] \quad ⑩$$

Note that $\|\mu\|_2^2 \cos^2\theta$ is a constant positive number under the constraint of ⑩. For any constant positive number C, the function $f(x) = 2\log\left(x + \frac{1}{x}\right) + \frac{C}{1+x^2}$ is monotonously decreasing in the interval of $\left(0, \sqrt{\frac{C+\sqrt{C^2+4}}{2}}\right]$ and monotonously increasing in the interval of $\left[\sqrt{\frac{C+\sqrt{C^2+4}}{2}}, \infty\right)$, which means that if for some θ, β^* is the optimal solution to ⑩, it is necessary that for the same θ, β^* is the optimal solution to either of the following two problems:

$$\beta^+(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\max} \beta^T\Lambda\beta \quad (11)$$

$$\beta^-(\theta) = \underset{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}{\arg\min} \beta^T\Lambda\beta \quad (12)$$

Following Lagrange multiplier method as before, we could obtain that the solutions have the same proportional structure.

5. Connection between One Dimensional Solutions to Zero-Mean Cases and Non-Zero-Mean Cases

A generic result is that in non-zero mean cases, the optimal solution $A \in \mathbb{R}^{1\times n}$ should satisfy:

$$(\Sigma + l_1 I_n)A^T = -\frac{l_2}{2\|\mu\|_2}\mu$$

for some l_1, l_2, whereas in zero-mean cases, the condition is reduced to:

$$(\Sigma + l_1 I_n)A^T = 0$$

This is consistent with our previous findings that the optimal solution A should be Σ's eigenvectors in zero-mean cases for KL divergence, symmetric KL divergence, and Hellinger distance.