Maximizing Hellinger Distance after Linear Dimensionality Reduction

Sihui Wang

Problem:

Suppose $X, Y \in \mathbb{R}^n$, $X \sim \mathcal{N}(0, I_n)$, $Y \sim \mathcal{N}(0, \Sigma)$, $A \in \mathbb{R}^{r \times n}$, then we have $X_\mathrm{A} = AX \sim \mathcal{N}(0, AA^T)$ and $Y_A \sim \mathcal{N}(0, A\Sigma A^T)$. Assume $Z = A\Sigma A^T$, $h_{X_A}(x), h_{Y_A}(x)$ are probability density functions of X_A, Y_A, respectively.

Find A to maximize the following:

$$D_H\left(h_{X_A}, h_{Y_A}\right) = \frac{1}{\sqrt{2}} \sqrt{\int \left(\sqrt{h_{X_A}(x)} - \sqrt{h_{Y_A}(x)}\right)^2 dx}$$

Solution:

1.Hellinger Distance under Gaussian Distribution:

Hellinger distance has closed form for Gaussian Distributions. According to [1], [2], the Hellinger distance between the distribution of $h_{X_A}(x)$ and the distribution of $h_{Y_A}(x)$ is:

$$D_H\left(h_{X_A}, h_{Y_A}\right) = \sqrt{1 - \frac{|AA^T|^{\frac{1}{4}} |A\Sigma A^T|^{\frac{1}{4}}}{\left|\frac{AA^T + A\Sigma A^T}{2}\right|^{\frac{1}{2}}}}$$

2.Formulation of the Problem:

Our problem is to find $A \in \mathbb{R}^{r \times n}$, so that:

$$D_H\left(h_{X_A}, h_{Y_A}\right) = \sqrt{1 - \frac{|AA^T|^{\frac{1}{4}} |A\Sigma A^T|^{\frac{1}{4}}}{\left|\frac{AA^T + A\Sigma A^T}{2}\right|^{\frac{1}{2}}}}$$

reaches its maxima.

3.Linear invariant of Hellinger Distance:

We can prove that Hellinger distance $D_H\left(h_{X_A}, h_{Y_A}\right)$ is invariant under linear transformations on the row vectors of A. To make it brief, we say that $D_H\left(h_{X_A}, h_{Y_A}\right)$ is linear invariant.

To prove $D_H\left(h_{X_A}, h_{Y_A}\right)$ is linear invariant, we can in turn prove $\frac{|AA^T|^{\frac{1}{4}} |A\Sigma A^T|^{\frac{1}{4}}}{\left|\frac{AA^T + A\Sigma A^T}{2}\right|^{\frac{1}{2}}}$ is linear invariant. To make it simpler, we can prove $d\left(h_{X_A}, h_{Y_A}\right) = \frac{|AA^T + A\Sigma A^T|^2}{|AA^T||A\Sigma A^T|}$ is linear invariant.

All linear transformations on the row vectors of A can be decomposed into 3 kinds of elementary components: a) multiplying the i-th row by k times; b) adding the j-th row to the i-th row; c) switching the i-th row and the j-th row. Operation a) can be represented by A's left multiplying $U_k(i) \in \mathbb{R}^{r \times r}$, operation b) can be represented by A's left multiplying $T(i,j) \in \mathbb{R}^{r \times r}$, and operation c) can be represented by A's left multiplying $S(i,j) \in \mathbb{R}^{r \times r}$:

$$U_k(i) = \begin{array}{cc} \begin{pmatrix} 1 &&&&&& \\ & \ddots &&&&& \\ && 1 &&&& \\ &&& k &&& \\ &&&& 1 && \\ &&&&& \ddots & \\ &&&&&& 1 \end{pmatrix} & \begin{matrix} \\ \\ \\ \leftarrow\ \ i-th\ \ row \\ \\ \\ \\ \end{matrix} \\ \begin{matrix} \uparrow \\ i-th\ \ column \end{matrix} & \end{array}$$

$$T(i,j) = \begin{array}{cc} \begin{pmatrix} 1 &&&&&& \\ & \ddots &&&&& \\ && 1 && 1 && \\ &&& \ddots &&& \\ &&&& 1 && \\ &&&&& \ddots & \\ &&&&&& 1 \end{pmatrix} & \begin{matrix} \\ \\ \leftarrow\ \ i-th\ \ row \\ \\ \\ \\ \\ \end{matrix} \\ \begin{matrix} \uparrow \\ j-th\ \ column \end{matrix} & \end{array}$$

$$S(i,j) = \begin{array}{cc} \begin{pmatrix} 1 &&&&&& \\ & \ddots &&&&& \\ && 0 & \cdots & 1 && \\ && \vdots & \ddots & \vdots && \\ && 1 & \cdots & 0 && \\ &&&&& \ddots & \\ &&&&&& 1 \end{pmatrix} & \begin{matrix} \\ \\ \leftarrow\ \ i-th\ \ row \\ \\ \\ \\ \\ \end{matrix} \\ \begin{matrix} \uparrow \\ j-th\ \ column \end{matrix} & \end{array}$$

Hence, to prove that $D_H\left(h_{X_A}, h_{Y_A}\right)$ is linear invariant, we just need to prove $d\left(h_{X_A}, h_{Y_A}\right) = d\left(h_{X_{\tilde{A}}}, h_{Y_{\tilde{A}}}\right)$ under the transformations of $\tilde{A} = U_k(i)A$, $\tilde{A} = T(i,j)A$, and $\tilde{A} = S(i,j)A$.

This is true, since for any invertible matrix $M \in \mathbb{R}^{r\times r}$, $\tilde{A} = MA$, we have:

$$d\left(h_{X_{\tilde{A}}}, h_{Y_{\tilde{A}}}\right) = \frac{|MAA^TM^T + MA\Sigma A^TM^T|^2}{|MAA^TM^T||MA\Sigma A^TM^T|} = \frac{|M|^2|AA^T + A\Sigma A^T|^2|M^T|^2}{|M||AA^T||M^T||M||A\Sigma A^T||M^T|} = \frac{|AA^T + A\Sigma A^T|^2}{|AA^T||A\Sigma A^T|}$$
$$= d(h_{X_A}, h_{Y_A})$$

Obviously, $U_k(i) \in \mathbb{R}^{r\times r}$, $T(i,j) \in \mathbb{R}^{r\times r}$, and $S(i,j) \in \mathbb{R}^{r\times r}$ are invertible matrices, so $d(h_{X_{\tilde{A}}}, h_{Y_{\tilde{A}}})$ is invariant under all linear transformations on the row vectors of *A*, which in turn proves that Hellinger distance is invariant under all linear transformations on the row vectors of *A*.

The implications of Hellinger distance's linear invariant.

As long as Hellinger distance is linear invariant, we can always obtain a matrix *A* whose row vectors are orthonormal vectors. That is to say, without any change to the maxima of Hellinger distance, we can always obtain a semi-orthogonal matrix, $A \in \mathbb{R}^{r\times n}$, satisfying $AA^T = I_r$.

4.Problem Reduction:

First, we can simplify our objective function.

To maximize $D_H\left(h_{X_A}, h_{Y_A}\right) = \sqrt{1 - \frac{|AA^T|^{\frac{1}{4}}|A\Sigma A^T|^{\frac{1}{4}}}{\left|\frac{AA^T + A\Sigma A^T}{2}\right|^{\frac{1}{2}}}}$ is to maximize $D_H^2\left(h_{X_A}, h_{Y_A}\right) = 1 -$

$\frac{|AA^T|^{\frac{1}{4}}|A\Sigma A^T|^{\frac{1}{4}}}{\left|\frac{AA^T+A\Sigma A^T}{2}\right|^{\frac{1}{2}}}$, which is to minimize $1 - D_H^2(h_{X_A}, h_{Y_A}) = \frac{|AA^T|^{\frac{1}{4}}|A\Sigma A^T|^{\frac{1}{4}}}{\left|\frac{AA^T+A\Sigma A^T}{2}\right|^{\frac{1}{2}}}$. This problem can in turn be converted to the problem of maximizing $\frac{1}{1-D_H^2(h_{X_A},h_{Y_A})} = \frac{\left|\frac{AA^T+A\Sigma A^T}{2}\right|^{\frac{1}{2}}}{|AA^T|^{\frac{1}{4}}|A\Sigma A^T|^{\frac{1}{4}}}$, which is equivalent to the problem of maximizing $\frac{|AA^T+A\Sigma A^T|^2}{|AA^T||A\Sigma A^T|}$.

Second, we consider the constraint. In section 3 we have proved that we can always assume that $AA^T = I_r$, so the objective function can be further simplified as $\frac{|I_r+A\Sigma A^T|^2}{|A\Sigma A^T|}$.

Hence, we can formulate the simplified optimization problem:

$$\max \frac{|I_r+A\Sigma A^T|^2}{|A\Sigma A^T|}$$
$$\text{s.t. } AA^T = I_r$$

5. Eigenvalue decomposition:

We know that the determinant of A is the product of all the eigenvalues of A. Hence, to maximize $\frac{|I_r+A\Sigma A^T|^2}{|A\Sigma A^T|}$, we got to analyze all the eigenvalues of $I_r + A\Sigma A^T$ and $Z = A\Sigma A^T$.

Suppose $\text{eig}(Z) = \gamma_1, \dots, \gamma_r$, then by eigenvalue decomposition we have an orthonormal basis in $\mathbb{R}^n$, $v_1, \dots, v_r$, satisfying $Z = \sum_{i=1}^r \gamma_i v_i v_i^T$. Since $I_r = \sum_{i=1}^r v_i v_i^T$, we have $I_r + A\Sigma A^T = \sum_{i=1}^r (\gamma_i + 1) v_i v_i^T$ which means that the eigenvalues of $I_r + A\Sigma A^T$ are $\gamma_1 + 1, \dots, \gamma_r + 1$.

Hence, we have:

$$\frac{|I_r + A\Sigma A^T|^2}{|A\Sigma A^T|} = \prod_{i=1}^{r} \frac{(\gamma_i + 1)^2}{\gamma_i}$$

So, to maximize the objective function $\frac{|I_r+A\Sigma A^T|^2}{|A\Sigma A^T|}$, we just need to individually maximize each factor, $\frac{(\gamma_i+1)^2}{\gamma_i}$.

6. Cauchy Interlacing Theorem:

Another crucial step is to establish the connection between the eigenvalues of Σ and the eigenvalues of $A\Sigma A^T$. This is done by applying Cauchy's interlacing theorem [3][4][5].

Suppose $\text{eig}(Z) = \gamma_1, \dots, \gamma_r, \gamma_1 \geq \dots \geq \gamma_r$ and $\text{eig}(\Sigma) = \lambda_1, \dots, \lambda_n, \lambda_1 \geq \dots \geq \lambda_n$, then according to Cauchy's interlacing theorem, we have:

$$\lambda_{n-r+i} \leq \gamma_i \leq \lambda_i \ \ (1 \leq i \leq r)$$

By Cauchy's interlacing theorem we established the lower bound and upper bound of $\gamma_i (1 \leq i \leq r)$. Next, we can use these results to find the maxima of $\frac{(\gamma_i+1)^2}{\gamma_i}$ and $\frac{|I_r+A\Sigma A^T|^2}{|A\Sigma A^T|} = \prod_{i=1}^r \frac{(\gamma_i+1)^2}{\gamma_i}$.

7. Results:

First, we make some observations.

a) Our aim is to maximize $\frac{|I_r+A\Sigma A^T|^2}{|A\Sigma A^T|}$, which is equivalent to maximize each of $\frac{(\gamma_i+1)^2}{\gamma_i}$. The latter is equivalent to maximize each of $\gamma_i + \frac{1}{\gamma_i}$.

b) $\gamma_i + \frac{1}{\gamma_i}$ is monotonously decreasing in the interval of $(0,1]$ and monotonously increasing in the interval of $[1,\infty)$. Besides, $\gamma_i + \frac{1}{\gamma_i}$ is convex when $\gamma_i > 0$. Either by its monotonicity or by its convexity we can prove that $\gamma_i + \frac{1}{\gamma_i}$ yields it maximum either at the upper bound of γ_i, λ_i, or at the lower bound of γ_i, λ_{n-r+i}.

c) As long as we construct the matrix *A* by the corresponding eigenvectors, for each $1 \le i \le r$ we can obtain $\gamma_i = \lambda_{n-r+i}$ or $\gamma_i = \lambda_i$. This observation means that each γ_i can achieve its theoretical upper bound and lower bound in reality.

Based on the above observations, to maximize $\frac{|I_r + A\Sigma A^T|^2}{|A\Sigma A^T|}$, we need to let each $\gamma_i = \lambda_i$ or $\gamma_i = \lambda_{n-r+i}$. That is to say, in order to maximize Hellinger distance, we need to choose *r* elements from $\lambda_i(1 \le i \le n)$ and assign them to $\gamma_i(1 \le i \le r)$.

So, in order to maximize Hellinger distance, we need to select the *r* elements from $\lambda_i(1 \le i \le n)$ who have the largest value of $\lambda_i + \frac{1}{\lambda_i}$. This is the short version of the results.

The long version of the results are as follows:

Case 1: Suppose $\lambda_1 \ge \cdots \ge \lambda_n \ge 1$, then $D_H\left(h_{X_A}, h_{Y_A}\right)$ yields its maximum by taking $\gamma_1 = \lambda_1,\ \gamma_2 = \lambda_2, \ldots, \gamma_r = \lambda_r$.

Case 2: Suppose $1 \ge \lambda_1 \ge \cdots \ge \lambda_n$, then $D_H\left(h_{X_A}, h_{Y_A}\right)$ yields its maximum by taking $\gamma_1 = \lambda_{n-r+1},\ \gamma_2 = \lambda_{n-r+2}, \ldots, \gamma_r = \lambda_n$.

Case 3: Suppose that there are both eigenvalues that are greater than 1 and eigenvalues that are less than 1. From the above discussions we have learned that $\gamma_i + \frac{1}{\gamma_i}$ yields its maximum either at the upper bound of γ_i, λ_i, or at the lower bound of γ_i, λ_{n-r+i}. So we can decide each γ_i's value by comparison of $\lambda_i + \frac{1}{\lambda_i}$ and $\lambda_{n-r+i} + \frac{1}{\lambda_{n-r+i}}$. If $\lambda_i + \frac{1}{\lambda_i} > \lambda_{n-r+i} + \frac{1}{\lambda_{n-r+i}}$, then we let $\gamma_i = \lambda_i$, otherwise we let $\gamma_i = \lambda_{n-r+i}$. This will suffice to obtain the maxima of Hellinger distance.

In practice, we usually don't need to make all the *r* comparisons. For example, if $\lambda_1\lambda_{n-r+1} < 1$ (which means that $\lambda_1 + \frac{1}{\lambda_1} < \lambda_{n-r+1} + \frac{1}{\lambda_{n-r+1}}$), then we let $\gamma_1 = \lambda_{n-r+1}, \gamma_2 = \lambda_{n-r+2}, \ldots, \gamma_r = \lambda_n$. We obtain the result by making only one comparison. In general, if $\lambda_i\lambda_{n-r+i} > 1$ and $\lambda_{i+1}\lambda_{n-r+i+1} < 1$, then we let $\gamma_1 = \lambda_1, \ldots, \gamma_i = \lambda_i$ and $\gamma_{i+1} = \lambda_{n-r+i+1}, \ldots, \gamma_r = \lambda_n$. Hence, we obtain the maxima of Hellinger distance by making $i+1$ comparisons. This method will equivalently get *r* eigenvalues of Σ with largest values of $\lambda_i + \frac{1}{\lambda_i}(1 \le i \le n)$, so the two versions of the results are basically the same.

8. Construction of *A*:

From the discussions in section 7 we have learned that we can construct *A* by the corresponding eigenvectors.

Specifically, in Case 1, we construct *A* by $A = \begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{pmatrix}$, in Case 2, we construct *A* by $A =$

$\begin{pmatrix} v_{n-r+1}^T \\ v_{n-r+2}^T \\ \vdots \\ v_n^T \end{pmatrix}$, in Case 3, if we choose $\gamma_1 = \lambda_1, \dots, \gamma_i = \lambda_i$ and $\gamma_{i+1} = \lambda_{n-r+i+1}, \dots, \gamma_r = \lambda_n$, then we construct A by $A = \begin{pmatrix} v_1^T \\ \vdots \\ v_i^T \\ v_{n-r+i+1}^T \\ \vdots \\ v_n^T \end{pmatrix}$. This will suffice to obtain the maxima of Hellinger distance.

9. Matrix A in general forms:

In section 3 we have learned that if Hellinger distance reaches its maxima by taking $A = A_0$, then Hellinger distance will reach the maxima by taking $A = MA_0$ provided $M \in \mathbb{R}^{\mathrm{r}\times\mathrm{r}}$ is any invertible matrix. From this we conclude that Hellinger distance will reach the maxima as long as A's row vectors span the same linear space as the one spanned by $v_1, \dots, v_r$ which correspond to the r-largest items in $\lambda_1 + \frac{1}{\lambda_1}, \lambda_2 + \frac{1}{\lambda_2}, \dots, \lambda_n + \frac{1}{\lambda_n}$.

Reference:

[1] L. Pardo, Statistical Inference Based on Divergence Measures, pp.45

[2] L. Devrove, A. Mehrabian, T. Reddad, The total variation distance between high-dimensional Gaussians.

[3] https://en.wikipedia.org/wiki/Poincar%C3%A9_separation_theorem

[4] https://en.wikipedia.org/wiki/Min-max_theorem#Cauchy_interlacing_theorem

[5] R. Bhatia, Matrix Analysis, pp. 59, Corollary III.1.5