Observations for Maximization of *f*-divergences: Optimization Perspective

Sihui Wang

1. Calculations of the gradients of the matrix functions:

1) Calculation of $\frac{\partial \log|A\Sigma A^T|}{\partial A}$:

What we want is $\frac{\partial \log|A\Sigma A^T|}{\partial A}$. Since $\frac{\partial \log|A\Sigma A^T|}{\partial A} = \frac{\partial \log|A\Sigma A^T|}{\partial |A\Sigma A^T|} \cdot \frac{\partial |A\Sigma A^T|}{\partial A} = \frac{1}{|A\Sigma A^T|} \frac{\partial |A\Sigma A^T|}{\partial A}$, the core task is to compute $\frac{\partial |A\Sigma A^T|}{\partial A}$.

So, let's find out what impact a small perturbation on the element at *i*-th row, *j*-th column of *A* will have on $|A\Sigma A^T|$.

We have:

$$\left(A + tE_{ij}\right)\Sigma\left(A + tE_{ij}\right)^T = A\Sigma A^T + tA\Sigma E_{ij}^T + tE_{ij}\Sigma A^T + t^2 E_{ij}\Sigma E_{ij}^T$$

$$= A\Sigma A^T + t\begin{pmatrix} 0 & \underset{\uparrow\ \text{i-th column}}{col_j(A\Sigma)} & 0 \end{pmatrix} + t\begin{pmatrix} 0 \\ row_j(\Sigma A^T) \\ 0 \end{pmatrix} \nwarrow_{\text{i-th row}} + t^2 E_{ij}\Sigma E_{ij}^T$$

Assume $A\Sigma = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{r1} & \cdots & y_{rn} \end{pmatrix}$, also notice that Σ is symmetric, so $row_j(\Sigma A^T) = \left[col_j(A\Sigma)\right]^T$, and we have:

$$\left(A + tE_{ij}\right)\Sigma\left(A + tE_{ij}\right)^T = A\Sigma A^T + t\begin{array}{c} \text{i-th column} \\ \left(\begin{array}{ccccc} 0 & \cdots & y_{1j} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ y_{1j} & \cdots & 2y_{jj} & \cdots & y_{rj} \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & y_{rj} & \cdots & 0 \end{array}\right) \end{array} \begin{array}{c} \\ \\ + O(t^2) \\ \\ \text{i-th row} \end{array}$$

Suppose $\left|\left(A + tE_{ij}\right)\Sigma\left(A + tE_{ij}\right)^T\right| = C_0 + C_1 t + O(t^2)$. Now we want to decide C_0 and C_1.

Recalling the formula for calculation of determinants:

$$|D| = \sum (-1)^{\tau(i_1 i_2 i_3 \dots i_n)} d_{1i_1} d_{2i_2} d_{3i_3} \dots d_{ni_n}$$

So, each time we choose the i_k-th element in the *k*-th row of $D (1 \le k \le n)$ and make sure that $i_1, \dots, i_n$ are mutually different. We multiply these factors together with a coefficient $+1$ or -1 to get one term in $|D|$. Taking the sum of all those terms, we get the determinant of *D*.

So, C_0 is the sum of those terms in $\left(A + tE_{ij}\right)\Sigma\left(A + tE_{ij}\right)^T$ that are free from *t*. It can be deduced that in fact $C_0 = |A\Sigma A^T|$.

Then let us find out what C_1 is. For example, if we choose y_{1j} at the 1-st row, *i*-th column of $\left(A+tE_{ij}\right)\Sigma\left(\mathrm{A}+tE_{ij}\right)^{\mathrm{T}}$, then, to make a term of order *t*, all the remaining factors should be outside 1-st row or *i*-th column, and they should be free from *t*. It can be deduced that the sum of all those terms is Z_{1i}. Here, $Z=A\Sigma A^T$ and Z_{ij} is the algebraic cofactor of *Z*.

So, we find out that in $\mathrm{C}_1 t$ there is a term $y_{1j}Z_{1i}$ as one of the components. Following the same deduction, and taking the fact that Z is symmetric into consideration, eventually we have:

$$\mathrm{C}_1 t=2\mathrm{row}_i(Z^{*T})col_j(A\Sigma)t$$

Where $Z^*=\begin{pmatrix} Z_{11} & \cdots & Z_{r1} \\ \vdots & \ddots & \vdots \\ Z_{1r} & \cdots & Z_{rr}\end{pmatrix}$ is the adjugate matrix of *Z*.

To compute the derivate of 1-st order, we can disregard the $\mathrm{O}(\mathrm{t}^2)$ terms in $|\left(A+tE_{ij}\right)\Sigma\left(\mathrm{A}+tE_{ij}\right)^{\mathrm{T}}|$. Hence we have:

$$\frac{\partial\left|\mathrm{A\Sigma A^T}\right|}{\partial a_{ij}}=\lim_{t\to 0}\frac{\left|\left(A+tE_{ij}\right)\Sigma\left(\mathrm{A}+tE_{ij}\right)^{\mathrm{T}}\right|-|A\Sigma A^T|}{t}=\lim_{t\to 0}\frac{C_1 t+O(t^2)}{t}=C_1$$

$$=2row_i(Z^{*T})col_j(A\Sigma)$$

So we have:

$$\frac{\partial\left|\mathrm{A\Sigma A^T}\right|}{\partial A}=2Z^{*T}A\Sigma$$

As $\mathrm{Z}^{-1}=\frac{1}{|Z|}Z^*$, we have:

$$\frac{\partial\left|\mathrm{A\Sigma A^T}\right|}{\partial A}=2|Z|Z^{-T}A\Sigma$$

Since $Z=A\Sigma A^T$ is symmetric, we have $Z^{-1}=Z^{-T}$ and:

$$\frac{\partial\left|\mathrm{A\Sigma A^T}\right|}{\partial A}=2\left|\mathrm{A\Sigma A^T}\right|(A\Sigma A^T)^{-1}A\Sigma$$

So finally we have:

$$\frac{\partial\log\left|\mathrm{A\Sigma A^T}\right|}{\partial \mathrm{A}}=\frac{\partial\log\left|\mathrm{A\Sigma A^T}\right|}{\partial\left|\mathrm{A\Sigma A^T}\right|}\cdot\frac{\partial\left|\mathrm{A\Sigma A^T}\right|}{\partial A}=\frac{1}{\left|\mathrm{A\Sigma A^T}\right|}\frac{\partial\left|\mathrm{A\Sigma A^T}\right|}{\partial A}=2(A\Sigma A^T)^{-1}A\Sigma$$

2) Calculation of $\frac{\partial\mathrm{tr}[(A\Sigma_1 A^T)^{-1}(\mathrm{A}\Sigma_2\mathrm{A^T})]}{\partial\mathrm{A}}$:

We can follow similar ideas to calculate $\mathrm{tr}[\left(\left(A+tE_{ij}\right)\Sigma_1\left(A+tE_{ij}\right)^T\right)^{-1}\left(A+tE_{ij}\right)\Sigma_2\left(A+\mathrm{t}E_{ij}\right)^{\mathrm{T}}]$.

$$\left(\left(A+tE_{ij}\right)\Sigma_1\left(A+tE_{ij}\right)^T\right)^{-1}=\left[A\Sigma_1 A^T+t\left(A\Sigma_1 E_{ij}^T+E_{ij}\Sigma_1 A^T\right)+\mathrm{O}(t^2)\right]^{-1}$$

According to theorems in functional analysis, if *T* is invertible and $||\Delta T||$ is sufficiently small, then $S=T+\Delta T$ is also invertible:

$$S^{-1}=\sum_{k=0}^{\infty}(-1)^k T^{-1}(T^{-1}\Delta T)^k$$

So, if *t* is sufficiently small, then $\left(A+tE_{ij}\right)\Sigma_1\left(A+tE_{ij}\right)^T$ is invertible. In our case, $T=$

$A\Sigma_1 A^T$ and $\Delta T = t\left(A\Sigma_1 E_{ij}^T + E_{ij}\Sigma_1 A^T\right) + O(t^2)$, hence we have:

$$\left(\left(A + tE_{ij}\right)\Sigma_1\left(A + tE_{ij}\right)^T\right)^{-1} = (A\Sigma_1 A^T)^{-1} - t(A\Sigma_1 A^T)^{-2}\left(A\Sigma_1 E_{ij}^T + E_{ij}\Sigma_1 A^T\right) + O(t^2)$$

$$\left(A + tE_{ij}\right)\Sigma_2\left(A + tE_{ij}\right)^T = A\Sigma_2 A^T + t\left(A\Sigma_2 E_{ij}^T + E_{ij}\Sigma_2 A^T\right) + O(t^2)$$

So we have:

$$\begin{aligned}\mathrm{tr}&\left[\left(\left(A + tE_{ij}\right)\Sigma_1\left(A + tE_{ij}\right)^T\right)^{-1}\left(A + tE_{ij}\right)\Sigma_2\left(A + tE_{ij}\right)^T\right]\\ &= tr[(A\Sigma_1 A^T)^{-1} A\Sigma_2 A^T] + t\cdot tr\left[(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T\right] + t\\ &\cdot tr\left[(A\Sigma_1 A^T)^{-1} E_{ij}\Sigma_2 A^T\right] - t\cdot tr\left[(A\Sigma_1 A^T)^{-2}\left(A\Sigma_1 E_{ij}^T + E_{ij}\Sigma_1 A^T\right)A\Sigma_2 A^T\right]\\ &+ O(t^2)\end{aligned}$$

As:

$$\frac{\partial \mathrm{tr}[(A\Sigma_1 A^T)^{-1}(A\Sigma_2 A^T)]}{\partial a_{ij}} = \lim_{t\to 0}\frac{\mathrm{tr}\left[\left(\left(A+tE_{ij}\right)\Sigma_1\left(A+tE_{ij}\right)^T\right)^{-1}\left(A+tE_{ij}\right)\Sigma_2\left(A+tE_{ij}\right)^T\right] - \mathrm{tr}[(A\Sigma_1 A^T)^{-1}(A\Sigma_2 A^T)]}{t},$$

we can disregard all the terms in $O(t^2)$, so we have:

$$\begin{aligned}\frac{\partial \mathrm{tr}[(A\Sigma_1 A^T)^{-1}(A\Sigma_2 A^T)]}{\partial a_{ij}}&\\ &= tr\left[(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T\right] + tr\left[(A\Sigma_1 A^T)^{-1} E_{ij}\Sigma_2 A^T\right]\\ &- tr\left[(A\Sigma_1 A^T)^{-2}\left(A\Sigma_1 E_{ij}^T + E_{ij}\Sigma_1 A^T\right)A\Sigma_2 A^T\right]\end{aligned}$$

First, we compute $tr\left[(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T\right]$.

Notice that

$$(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T = (A\Sigma_1 A^T)^{-1}\begin{array}{c}\begin{array}{ccc} & \text{i-th column} & \\ & \downarrow & \end{array}\\ \begin{pmatrix}0 & col_j(A\Sigma_2) & 0\end{pmatrix}\end{array}$$

So $\mathrm{tr}[(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T]$ is in fact the element at *i*-th row and *i*-th column of $(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T$, which is:

$$\mathrm{tr}\left[(A\Sigma_1 A^T)^{-1} A\Sigma_2 E_{ij}^T\right] = row_i(A\Sigma_1 A^T)^{-1}\cdot col_j(A\Sigma_2)$$

Similarly, we can obtain:

$$(A\Sigma_1 A^T)^{-1} E_{ij}\Sigma_2 A^T = (A\Sigma_1 A^T)^{-1}\begin{array}{cl}\begin{pmatrix}0\\ row_j(\Sigma_2 A^T)\\ 0\end{pmatrix} & \begin{array}{l}\\ \leftarrow \text{i-th row}\\ \\ \end{array}\end{array}$$

Since $tr(AB) = tr(BA)$, we have:

$$\mathrm{tr}\left[(A\Sigma_1 A^T)^{-1}\begin{pmatrix}0\\ row_j(\Sigma_2 A^T)\\ 0\end{pmatrix}\right] = tr[\begin{pmatrix}0\\ row_j(\Sigma_2 A^T)\\ 0\end{pmatrix}(A\Sigma_1 A^T)^{-1}]$$

$tr[\begin{pmatrix}0\\ row_j(\Sigma_2 A^T)\\ 0\end{pmatrix}(A\Sigma_1 A^T)^{-1}]$ is in fact the element at *i*-th row and *i*-th column of $\begin{pmatrix}0\\ row_j(\Sigma_2 A^T)\\ 0\end{pmatrix}(A\Sigma_1 A^T)^{-1}$. So we have:

$$tr\left[(A\Sigma_1 A^T)^{-1} E_{ij}\Sigma_2 A^T\right] = row_j(\Sigma_2 A^T) col_i(A\Sigma_1 A^T)^{-1}$$

So, $tr\left[(A\Sigma_1 A^T)^{-1} E_{ij}\Sigma_2 A^T\right]$ is the element at *j*-th row and *i*-th column of $\Sigma_2 A^T(A\Sigma_1 A^T)^{-1}$, which is the element at *i*-th row and *j*-th column of $[(A\Sigma_1 A^T)^{-1}]^T(\Sigma_2 A^T)^T$. So:

$$tr\left[(A\Sigma_1 A^T)^{-1} E_{ij}\Sigma_2 A^T\right] = row_i(A\Sigma_1 A^T)^{-1}\cdot col_j(A\Sigma_2)$$

For $tr\left[(A\Sigma_1 A^T)^{-2}\left(A\Sigma_1 E_{ij}^T + E_{ij}\Sigma_1 A^T\right)A\Sigma_2 A^T\right]$, we break it into two terms, $\mathrm{tr}[(A\Sigma_1 A^T)^{-2}A\Sigma_1 E_{ij}^T A\Sigma_2 A^T]$ and $\mathrm{tr}[(A\Sigma_1 A^T)^{-2}E_{ij}\Sigma_1 A^T A\Sigma_2 A^T]$.

For the first part, $\mathrm{tr}[(A\Sigma_1 A^T)^{-2}A\Sigma_1 E_{ij}^T A\Sigma_2 A^T]$, according to $tr(AB) = tr(BA)$ we have $\mathrm{tr}\left[(A\Sigma_1 A^T)^{-2}A\Sigma_1 E_{ij}^T A\Sigma_2 A^T\right] = \mathrm{tr}[A\Sigma_2 A^T (A\Sigma_1 A^T)^{-2}A\Sigma_1 E_{ij}^T]$.

Recalling that in the computation of $\mathrm{tr}\left[(A\Sigma_1 A^T)^{-1}A\Sigma_2 E_{ij}^T\right] = row_i(A\Sigma_1 A^T)^{-1}\cdot col_j(A\Sigma_2)$, we have learned how to compute arbitrary expressions of $\mathrm{tr}\left[(\cdot)A\Sigma_2 E_{ij}^T\right]$. So here we have:

$\mathrm{tr}\left[(A\Sigma_1 A^T)^{-2}A\Sigma_1 E_{ij}^T A\Sigma_2 A^T\right] = \mathrm{tr}\left[A\Sigma_2 A^T(A\Sigma_1 A^T)^{-2}A\Sigma_1 E_{ij}^T\right] = \mathrm{row_i} A\Sigma_2 A^T (A\Sigma_1 A^T)^{-2}\cdot col_j A\Sigma_1$

For the second part, $\mathrm{tr}[(A\Sigma_1 A^T)^{-2}E_{ij}\Sigma_1 A^T A\Sigma_2 A^T]$, according to $tr(AB) = tr(BA)$ we have $\mathrm{tr}\left[(A\Sigma_1 A^T)^{-2}E_{ij}\Sigma_1 A^T A\Sigma_2 A^T\right] = \mathrm{tr}[A\Sigma_2 A^T(A\Sigma_1 A^T)^{-2}E_{ij}\Sigma_1 A^T]$.

Recalling that in the computation of $tr\left[(A\Sigma_1 A^T)^{-1}E_{ij}\Sigma_2 A^T\right] = row_i[(A\Sigma_1 A^T)^{-1}]^T\cdot col_j(\Sigma_2 A^T)^T$, we have learned how to compute arbitrary expressions of $\mathrm{tr}\left[(\cdot)E_{ij}\Sigma_2 A^T\right]$. So here we have:

$$\begin{aligned}\mathrm{tr}\left[(A\Sigma_1 A^T)^{-2}E_{ij}\Sigma_1 A^T A\Sigma_2 A^T\right] &= \mathrm{tr}\left[A\Sigma_2 A^T(A\Sigma_1 A^T)^{-2}E_{ij}\Sigma_1 A^T\right]\\ &= row_i(A\Sigma_1 A^T)^{-2}A\Sigma_2 A^T col_j A\Sigma_1\end{aligned}$$

Hence, we have completed the calculation of all components of $\frac{\partial \mathrm{tr}[(A\Sigma_1 A^T)^{-1}(A\Sigma_2 A^T)]}{\partial a_{ij}}$. Notice that each component can be expressed in the form of $\mathrm{row_i}(\cdot)\cdot col_j(\cdot)$, so we can obtain the following:

$$\frac{\partial \mathrm{tr}[(A\Sigma_1 A^T)^{-1}(A\Sigma_2 A^T)]}{\partial A} = 2(A\Sigma_1 A^T)^{-1}A\Sigma_2 - (A\Sigma_2 A^T)(A\Sigma_1 A^T)^{-2}A\Sigma_1 - (A\Sigma_1 A^T)^{-2}(A\Sigma_2 A^T)A\Sigma_1$$

The formulas given above are adequate for our discussions. However, for convenience we can give more formulas:

$$\frac{\partial log|A\Sigma A^T|}{\partial A} = 2(A\Sigma A^T)^{-1}A\Sigma$$

$$\frac{\partial log|AA^T|}{\partial A} = 2(AA^T)^{-1}A$$

$$\frac{\partial log|A|}{\partial A} = A^{-T}$$

$$\frac{\partial log|AB|}{\partial A} = (AB)^{-T}B^T$$

$$\frac{\partial log|AB|}{\partial B} = A^T(AB)^{-T}$$

$$\frac{\partial tr(A)}{\partial A} = I_n$$

$$\frac{\partial tr(A^{-1})}{\partial A} = (A^{-2})^T$$

$$\frac{\partial \mathrm{tr}(AB)}{\partial A} = B^T$$

$$\frac{\partial tr(A\Sigma A^T)}{\partial A} = 2A\Sigma$$

$$\frac{\partial tr[(A\Sigma A^T)^{-1}]}{\partial A} = -2(A\Sigma A^T)^{-2}A\Sigma$$

$$\frac{\partial tr[(A\Sigma A^T)^{-1}(AA^T)]}{\partial A} = 2(A\Sigma A^T)^{-1}A - AA^T(A\Sigma A^T)^{-2}A\Sigma - (A\Sigma A^T)^{-2}AA^TA\Sigma$$

$$\frac{\partial log|A\Sigma A^T + AA^T|}{\partial A} = 2(A\Sigma A^T + AA^T)^{-1}A(\Sigma + I_n)$$

$$\frac{\partial log|A\Sigma_1 A^T + A\Sigma_2 A^T|}{\partial A} = 2(A(\Sigma_1 + \Sigma_2)A^T)^{-1}A(\Sigma_1 + \Sigma_2)$$

$$\frac{\partial tr[(A\Sigma_1 A^T)^{-1}(A\Sigma_2 A^T)]}{\partial A} = 2(A\Sigma_1 A^T)^{-1}A\Sigma_2 - A\Sigma_2 A^T(A\Sigma_1 A^T)^{-2}A\Sigma_1 - (A\Sigma_1 A^T)^{-2}A\Sigma_2 A^T A\Sigma_1$$

2. Formulation of the optimization problem:

According to the formulas given above, we can calculate the gradients for KL-divergence, symmetric KL-divergence and Hellinger distance.

1) KL-divergence:

$$\nabla_A KL = \frac{1}{2}[\frac{\partial \log|A\Sigma A^T|}{\partial A} - \frac{\partial \log|AA^T|}{\partial A} + \frac{\partial tr[(A\Sigma A^T)^{-1}(AA^T)]}{\partial A}]$$

$$= (A\Sigma A^T)^{-1}A\Sigma - (AA^T)^{-1}A + (A\Sigma A^T)^{-1}A - \frac{1}{2}AA^T(A\Sigma A^T)^{-2}A\Sigma$$

$$-\frac{1}{2}(A\Sigma A^T)^{-2}AA^TA\Sigma$$

In our problem, we should find the global maxima where $\nabla_A KL = 0$. According to previous discussions, KL-divergence is invariant under invertible linear transformations of rank *r*, so for any global optimal solution A, we can always find another global optimal solution $\tilde{A}$ satisfying the constraint $\tilde{A}\widetilde{A^T} = I_r$, and $\tilde{A} = RA$, $R \in \mathbb{R}^{r\times r}$ is invertible. So, without loss of generality, we can constrain ourselves to the points where $\nabla_A KL = 0$ and $AA^T = I_r$.

In the above equation, we let $AA^T = I_r$ and obtain:

$$\nabla_A KL = (A\Sigma A^T)^{-1}A\Sigma - A + (A\Sigma A^T)^{-1}A - (A\Sigma A^T)^{-2}A\Sigma$$

Note that we should calculate the gradient first and then impose the constraint of $AA^T = I_r$. If we do it in reverse order, we will get a wrong answer.

From the above equation, we can deduce that $\nabla_A KL = 0$ is equivalent to:

$$[I_r - (A\Sigma A^T)^{-1}][(A\Sigma A^T)^{-1}A\Sigma - A] = 0$$

If $I_r - (A\Sigma A^T)^{-1} = 0$, then $D_{KL} = 0$. This is the global minima. So we need $(A\Sigma A^T)^{-1}A\Sigma - A = 0$ to obtain the maxima.

2) Symmetric KL-divergence:

$$D_{\text{Sym}} = \frac{1}{2}tr[(A\Sigma A^T)^{-1}(AA^T)] + \frac{1}{2}tr[(AA^T)^{-1}(A\Sigma A^T)] - r$$

So:

$$\nabla_A Sym = (A\Sigma A^T)^{-1}A - \frac{1}{2}AA^T(A\Sigma A^T)^{-2}A\Sigma - \frac{1}{2}(A\Sigma A^T)^{-2}AA^TA\Sigma + (AA^T)^{-1}A\Sigma$$

$$-\frac{1}{2}A\Sigma A^T(AA^T)^{-2}A - \frac{1}{2}(AA^T)^{-2}A\Sigma A^TA$$

Let $AA^T = I_r$, we have:

$$\nabla_A Sym = (A\Sigma A^T)^{-1}A - (A\Sigma A^T)^{-2}A\Sigma + A\Sigma - A\Sigma A^TA$$

So $\nabla_A Sym = 0$ is equivalent to:

$$[I_r - (A\Sigma A^T)^{-2}][(A\Sigma A^T)A - A\Sigma] = 0$$

Suppose $I_r - (A\Sigma A^T)^{-2} = 0$. Assume $(A\Sigma A^T)^{-1}$'s eigenvalues are $\lambda_1, \ldots, \lambda_r$, then

$(A\Sigma A^T)^{-2}$'s eigenvalues are $\lambda_1^2, \dots, \lambda_r^2$. Since $I_r - (A\Sigma A^T)^{-2} = 0$, we have $\lambda_1^2 = \cdots = \lambda_r^2 = 1$.

If Σ is positive definite and A is of rank r, then $A\Sigma A^T$ is also positive definite. So $\lambda_1, \dots, \lambda_r > 0$, which means that $\lambda_1 = \cdots = \lambda_r = 1$. Hence we have $A\Sigma A^T = I_r$ and $D_{Sym} = 0$. So in our problems $I_r - (A\Sigma A^T)^{-2}$ corresponds to the global minima. We need $(A\Sigma A^T)A - A\Sigma = A\Sigma(A^T A - I_n) = 0$ to obtain the maxima.

Note that $(A\Sigma A^T)^{-1}A\Sigma - A = 0$ is equivalent to $(A\Sigma A^T)A - A\Sigma = 0$, so basically KL-divergence and symmetric KL-divergence will simultaneously reach the maxima.

3) Hellinger distance:

From previous discussions, we have learned that maximizing Hellinger distance is equivalent to maximizing the following:

$$\frac{|A\Sigma A^T + AA^T|^2}{|A\Sigma A^T||AA^T|}$$

Which is equivalent to maximizing the following:

$$2\log|A\Sigma A^T + AA^T| - log|A\Sigma A^T| - log|AA^T|$$

According to the formulas given in section 1, we can calculate the gradient of the above expression. And we can deduce that $\nabla_A H = 0$ is equivalent to:

$$2(A\Sigma A^T + I_r)^{-1}A(\Sigma + I_n) - (A\Sigma A^T)^{-1}A\Sigma - A = 0$$

As each objective function has different gradient, we cannot analyze them in a generic way. In the following section, we will discuss the condition when $\nabla_A KL = 0$. As we have mentioned above, $\nabla_A Sym = 0$ is equivalent to $\nabla_A KL = 0$ so we don't have to have an additional discussion for symmetric KL-divergence.

3. The uniqueness of the linear subspace for KL-divergence and symmetric KL-divergence:

We can verify that $\nabla_A KL = 0$ when $A \in \mathbb{R}^{r\times n}$'s row vectors are constructed as *r* eigenvectors of Σ. According to the linear invariance, we can deduce that $\nabla_A KL = 0$ as long as $A \in \mathbb{R}^{r\times n}$'s row vectors span a linear subspace that is identical to the linear subspace that is spanned by some *r* eigenvectors of Σ. However, it is not obvious whether or not $\nabla_A KL \neq 0$ when *A*'s *r* linear independent row vectors cannot be expressed as linear combinations of certain *r* eigenvectors of Σ. Now we prove that it is true: except for the global minimal point, the following conditions are equivalent:

1) $\nabla_A KL = 0$;

2) $A \in \mathbb{R}^{r\times n}$'s row vectors span a linear subspace that is identical to the linear subspace that is spanned by some *r* eigenvectors of Σ.

This implies that, except for the global minimal point, all the points satisfying $\nabla_A KL = 0$ can be divided into $\binom{n}{r}$ equivalence classes, each corresponds to a linear subspace spanned by certain *r* eigenvectors of Σ. In these $\binom{n}{r}$ equivalence classes, we can always identify at least one equivalence class where D_{KL} reaches the maxima.

To prove this, first we have to answer the question: what does it mean by "*A*'s *r* linear independent row vectors cannot be expressed as linear combinations of certain *r* eigenvectors of Σ"?

We can explain this by an example. Suppose $v_1, \dots, v_n$ are Σ's eigenvectors which form an orthonormal basis of $\mathbb{R}^n$. If $A = \binom{v_1 + v_3}{v_2 + v_3}$, then we can say that *A*'s row vectors cannot be linear expressed by any 2 eigenvectors of Σ.

Roughly speaking, if $A \in \mathbb{R}^{r\times n}$'s row vectors "use" more than *r* symbols among $v_1, \dots, v_n$, then we say that A's row vectors cannot be linear expressed by any *r* eigenvectors of Σ.

Moreover, we observe that $\nabla_A KL$ is invariant under orthogonal transformations. So, if *A*'s row vectors "use" more than *r* symbols among $v_1, \dots, v_n$, then up to an orthogonal transformation, we can express *A* by:

$$A = \begin{pmatrix} v_1' \\ \vdots \\ v_{r-1}' \\ \sum_{i=r}^{n} \mu_i v_i' \end{pmatrix}$$

Where $\sum_{i=r}^{n} \mu_i^2 = 1$, and there are more than one non-zero elements among μ_i^2.

Hence, to prove our statement is to prove the following:

$(A\Sigma A^T)A - A\Sigma \neq 0$, $AA^T = I_r$ if and only if up to an orthogonal transformation, $A = \begin{pmatrix} v_1' \\ \vdots \\ v_{r-1}' \\ \sum_{i=r}^{n} \mu_i v_i' \end{pmatrix}$, $\sum_{i=r}^{n} \mu_i^2 = 1$ and there are more than one non-zero elements among μ_i^2.

We just need to verify that $A\Sigma A^T A \neq A\Sigma$. Suppose by eigenvalue decomposition we have $\Sigma = (v_1 \ \cdots \ v_n)\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}\begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix}$, then:

$$A\Sigma = \begin{pmatrix} \lambda_1 v_1' \\ \vdots \\ \lambda_{r-1} v_{r-1}' \\ \sum_{i=r}^{n} \lambda_i \mu_i v_i' \end{pmatrix}$$

$$A\Sigma A^T = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_{r-1} & \\ & & & \sum_{i=r}^{n} \lambda_i \mu_i^2 \end{pmatrix}$$

$$A\Sigma A^T A = \begin{pmatrix} \lambda_1 v_1' \\ \vdots \\ \lambda_{r-1} v_{r-1}' \\ \sum_{i=r}^{n} \lambda_i \mu_i^2 \cdot \sum_{j=r}^{n} \mu_j v_j' \end{pmatrix}$$

Suppose there is only one element, say, μ_r, that is non-zero among $\mu_i (r \le i \le n)$. Then by $\sum_{i=r}^{n} \mu_i^2 = 1$ we can deduce that $\mu_r^2 = 1$ and $\mu_i^2 = 0$ for any $i \neq r, r \le i \le n$. Then $\sum_{i=r}^{n} \lambda_i \mu_i v_i' = \sum_{i=r}^{n} \lambda_i \mu_i^2 \cdot \sum_{j=r}^{n} \mu_j v_j' = \lambda_r \mu_r v_r'$. In this case, we have $A\Sigma = A\Sigma A^T A$, that is, $\nabla_A KL = 0$.

However, if there are more than one non-zero elements among $\mu_i (r \le i \le n)$, then without loss of generality we denote them by $\mu_r, \dots, \mu_{r+k} (k \ge 1)$. We can also assume that the corresponding eigenvalues, $\lambda_r, \dots, \lambda_{r+k}$, satisfy $\lambda_r > \cdots > \lambda_{r+k}$. Then we argue that we must have $\lambda_r = \sum_{i=r}^{n} \lambda_i \mu_i^2$ if we want $A\Sigma = A\Sigma A^T A$. However, as there are more than one non-zero elements in μ_i^2 and $\lambda_r > \cdots > \lambda_{r+k}$, we have $\sum_{i=r}^{n} \lambda_i \mu_i^2 < \lambda_r$, so $A\Sigma \neq A\Sigma A^T A$ when there are more than one non-zero elements among $\mu_i (r \leq i \leq n)$.

Hence, we find the necessary and sufficient condition for $\nabla_A KL = 0$:

If $\Sigma \in \mathbb{R}^{n\times n}$ has n different eigenvalues, then $\nabla_A KL = 0$ if and only if either of the following conditions is met:

1) A is the global minimal point satisfying $A\Sigma A^T = I_r$;

2) $A \in \mathbb{R}^{r\times n}$'s row vectors are constructed as r eigenvectors of Σ;

3) $A \in \mathbb{R}^{r\times n}$'s row vectors are linear independent, and can be expressed as linear combinations of certain r eigenvectors of Σ.

Previously, we construct A by heuristic methods. Now we proved that previously we have identified all the possible optimal solutions. To conclude, despite that A has infinite number of choices, the choice of eigenvalues and the choice of the linear subspace in which A's row vectors serve as a basis, are quite unique. If all of Σ's eigenvalues are different, and the evaluation function doesn't get the same value at two different eigenvalues of Σ (for example, $\lambda = 2$ and $\lambda = \frac{1}{2}$ are considered the same by the evaluation function $\lambda + \frac{1}{\lambda}$), then the choice of Σ's eigenvalues, and the choice of the corresponding linear subspace that is able to maximize f-divergences, are unique. If we consider $A\Sigma A^T$ as a compression, then under mild assumptions this compression is unique.

4. A simplified, generic proof for uniqueness:

In the last section we have proved that the optimal solution for KL-divergence is, in some sense, unique. Our proof seems to depend on the particular form of D_{KL} and $\nabla_A KL$. Since each f-divergence has a different objective function, it seems that we should prove the uniqueness of optimal solution on a case-by-case basis. This is, however, not true. I realized that, we can prove the uniqueness in all cases (KL-divergence, symmetric KL-divergence, Hellinger distance, etc.) using a generic method.

The method is generalized from the proof in the last section.

Assume Σ has n different eigenvalues, $\lambda_1 > \cdots > \lambda_n$. Suppose by eigenvalue decomposition $\Sigma = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix}$, $v_1, \dots, v_n$ is an orthonormal basis of $\mathbb{R}^n$.

Now we suppose that the semi-orthogonal matrix A's r row vectors are not linear combinations of certain r eigenvectors of Σ.

Suppose A's row vectors "use" all the n symbols, $v_1, \dots, v_n$. then up to an orthogonal transformation we have:

$$A = \begin{pmatrix} v_1' \\ \vdots \\ v_{r-1}' \\ \sum_{i=r}^{n} \mu_i v_i' \end{pmatrix}$$

Where $\sum_{i=r}^{n} \mu_i^2 = 1$. Note that in the previous section we have obtained the following:

$$A\Sigma A^T = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_{r-1} & \\ & & & \sum_{i=r}^{n} \lambda_i \mu_i^2 \end{pmatrix}$$

Assume $A\Sigma A^T$'s eigenvalues are $\gamma_1 > \cdots > \gamma_r$. As A uses all the n symbols, we have $\mu_i \neq 0 (r \leq i \leq n)$ and $\lambda_\mathrm{r} > \sum_{i=r}^{n} \lambda_i \mu_i^2 > \lambda_n$. Hence we have:

$$\lambda_n < \gamma_r < \lambda_r$$

Similarly, up to an orthogonal transformation we can rewrite A as:

$$A = \begin{pmatrix} v_1' \\ \vdots \\ v_{r-2}' \\ \sum_{i=r-1}^{n-1} \mu_i v_i' \\ v_n' \end{pmatrix}$$

And we have:

$$A\Sigma A^T = \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_{r-2} & & \\ & & & \sum_{i=r-1}^{n-1} \lambda_i \mu_i^2 & \\ & & & & \lambda_n \end{pmatrix}$$

By the same deduction, we have:

$$\lambda_{n-1} < \gamma_{r-1} < \lambda_{r-1}$$

Repeat this argument for r times, we have:

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

when A "uses" all the n symbols in $v_1, \ldots, v_n$.

So what if A "uses" $n-k > r$ symbols in $v_1, \ldots, v_n$? In that case, we simply assume that A uses $v_{j_1}, \ldots, v_{j_{n-k}}$. Without loss of generality we assume the corresponding eigenvalues satisfy $\lambda_{j_1} > \cdots > \lambda_{j_{n-k}}$, then by the same deduction, we have:

$$\lambda_{j_{n-k-r+i}} < \gamma_i < \lambda_{j_i} (1 \leq i \leq r)$$

Because $\lambda_{j_1} > \cdots > \lambda_{j_{n-k}}$ are n-k items that are selected from $\lambda_1 > \cdots > \lambda_n$, we have:

$$\lambda_{i+k} \leq \lambda_{j_i} \leq \lambda_i (1 \leq i \leq n-k)$$

So we have:

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

Hence we proved that, if A's r row vectors are not linear combinations of certain r eigenvectors of Σ, then we have:

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

If A's r row vectors are linear combinations of certain r eigenvectors of Σ, say, $\lambda_{j_1} > \cdots > \lambda_{j_r}$, then we have:

$$\gamma_i = \lambda_{j_i} (1 \leq i \leq r)$$

To conclude, based on this method we can confirm that the optimal choice of A is, in some sense, unique. In fact, we used a different method to prove Cauchy's Interlacing theorem, and we proved that in Cauchy's Interlacing theorem, the equalities hold if and only if A's r row vectors are linear combinations of certain r eigenvectors of Σ.