Maximizing Divergence after Linear Dimensionality Reduction

Sihui Wang, Jan, 4th, 2020

Solution to Problem 1

For Gaussian random variable $X \in \mathbb{R}^n$, $X \sim \mathcal{N}(0, I)$, its probability density function (PDF), $f(\boldsymbol{x})$, is:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}}$$

in which $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ is a vector, $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$, and $\boldsymbol{x}'$ means the transpose of $\boldsymbol{x}$.

Similarly, for Gaussian random variable $Y \in \mathbb{R}^n$, $Y \sim \mathcal{N}(0, \Sigma)$, its PDF, $g(\boldsymbol{x})$, is:

$$g(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}}$$

By the definition of Kullback-Leiber (KL) divergence, we have:

$$D_{\mathrm{KL}}(f\|g) = \int_{\mathbb{R}^n} f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} d\boldsymbol{x} = \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} \log\left(|\Sigma|^{\frac{1}{2}} e^{\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}'\boldsymbol{x}}\right) d\boldsymbol{x}$$

hence:

$$D_{\mathrm{KL}}(f\|g) = \frac{1}{2}\log|\Sigma| \int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} d\boldsymbol{x} - \frac{1}{2}\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} \boldsymbol{x}'\boldsymbol{x} d\boldsymbol{x} + \frac{1}{2}\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} \boldsymbol{x}'\Sigma^{-1}\boldsymbol{x} d\boldsymbol{x}$$

So, to compute $D_{\mathrm{KL}}(f\|g)$, we just have to i) compute $\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} d\boldsymbol{x}$, ii) compute $\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} \boldsymbol{x}'\boldsymbol{x} d\boldsymbol{x}$, and iii) compute $\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} \boldsymbol{x}'\Sigma^{-1}\boldsymbol{x} d\boldsymbol{x}$.

i) Since $\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}}$ is the PDF of X, we simply have $\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} d\boldsymbol{x} = \int_{\mathbb{R}^n} f(\boldsymbol{x}) d\boldsymbol{x} = 1$.

ii) Since $X \sim \mathcal{N}(0, I)$, we have $\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}} \boldsymbol{x}'\boldsymbol{x} d\boldsymbol{x} = \mathbb{E}X'X = \mathbb{E}[(X - \mathbb{E}X)'(X - \mathbb{E}X)]$.

$X - \mathbb{E}X \in \mathbb{R}^{n\times 1}$, so $(X - \mathbb{E}X)'(X - \mathbb{E}X)$ is a scalar-valued random variable. We can consider $(X - \mathbb{E}X)'(X - \mathbb{E}X)$ as a 1×1 matrix, so:

$$\mathbb{E}[(X - \mathbb{E}X)'(X - \mathbb{E}X)] = \mathbb{E}[tr(X - \mathbb{E}X)'(X - \mathbb{E}X)]$$

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

$$\mathbb{E}[(X - \mathbb{E}X)'(X - \mathbb{E}X)] = \mathbb{E}[tr(X - \mathbb{E}X)(X - \mathbb{E}X)']$$

It is easy to see that $\mathbb{E}[tr(A)] = tr[\mathbb{E}(A)]$, for arbitrary $A \in \mathbb{R}^{n\times n}$, $\mathrm{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$:

$$\mathbb{E}[tr(A)] = \mathbb{E}\left[\sum_{i=1}^{n} a_{ii}\right] = \sum_{i=1}^{n} \mathbb{E}a_{ii} = tr[\mathbb{E}(A)]$$

Hence we have:

$$\mathbb{E}[(X-\mathbb{E}X)'(X-\mathbb{E}X)] = \mathbb{E}[tr(X-\mathbb{E}X)(X-\mathbb{E}X)'] = tr[\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)']$$

$\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)'$ is the covariance matrix of X, which means that $\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)' = I_{\mathrm{n}}$. Therefore, we have:

$$\mathbb{E}[(X-\mathbb{E}X)'(X-\mathbb{E}X)] = trI_n = n$$

Hence, we obtain that:

$$\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}x'x} \boldsymbol{x}'\boldsymbol{x}d\boldsymbol{x} = \mathbb{E}[(X-\mathbb{E}X)'(X-\mathbb{E}X)] = trI_n = n$$

iii) Similar to the calculation procedure in ii), we have:

$$\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}x'x} \boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}d\boldsymbol{x} = \mathbb{E}[X'\Sigma^{-1}X] = \mathbb{E}[tr(X'\Sigma^{-1}X)] = \mathbb{E}[tr(\Sigma^{-1}XX')]$$

$$\mathbb{E}[tr(\Sigma^{-1}XX')] = tr[\mathbb{E}(\Sigma^{-1}XX')] = tr[\Sigma^{-1}\mathbb{E}(XX')] = tr[\Sigma^{-1}I_n] = tr(\Sigma^{-1})$$

So

$$\int \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}x'x} \boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}d\boldsymbol{x} = tr(\Sigma^{-1})$$

Summing the results from i), ii), iii) together, we have:

$$D_{\mathrm{KL}}(f\|g) = \frac{1}{2}\log|\Sigma| - \frac{1}{2}n + \frac{1}{2}tr(\Sigma^{-1})$$

Solution to Problem 2

Since $X\sim\mathcal{N}(0, I_n)$, $Y\sim\mathcal{N}(0,\Sigma)$, $A\in\mathbb{R}^{r\times n}$, $r<n$, we have $X_A = AX\sim\mathcal{N}(0, AA')$, $Y_A = AY\sim\mathcal{N}(0, A\Sigma A')$. If Σ is non-singular and $\mathrm{rank}(A) = r$, then the PDF of X_A, $f_A(\boldsymbol{x})$, is:

$$f_A(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{r}{2}}|AA'|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'(AA')^{-1}\boldsymbol{x}}$$

The PDF of Y_A, $g_A(\boldsymbol{x})$, is:

$$g_A(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{r}{2}}|A\Sigma A'|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'(A\Sigma A')^{-1}\boldsymbol{x}}$$

So

$$D_{\mathrm{KL}}(f_A\|g_A) = \int f_A(\boldsymbol{x})\log\frac{f_A(\boldsymbol{x})}{g_A(\boldsymbol{x})}d\boldsymbol{x}$$

$$= \int \frac{1}{(2\pi)^{\frac{r}{2}}|AA'|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{x}'(AA')^{-1}\boldsymbol{x}}\log\left(\frac{|A\Sigma A'|^{\frac{1}{2}}}{|AA'|^{\frac{1}{2}}} e^{\frac{1}{2}\boldsymbol{x}'(A\Sigma A')^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}'(AA')^{-1}\boldsymbol{x}}\right)d\boldsymbol{x}$$

Therefore

$$D_{\mathrm{KL}}(f_A\|g_A) = \frac{1}{2}\log\frac{|A\Sigma A'|}{|AA'|}\int f_A(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int f_A(x)\boldsymbol{x}'(A\Sigma A')^{-1}\boldsymbol{x}d\boldsymbol{x} - \frac{1}{2}\int f_A(\boldsymbol{x})\boldsymbol{x}'(AA')^{-1}\boldsymbol{x}d\boldsymbol{x}$$

Following the same procedure like the solution to problem 1, we have:

$$\int_{\mathbb{R}^r} f_A(\boldsymbol{x})d\boldsymbol{x} = 1$$

$$\int f_A(x)\boldsymbol{x}'(A\Sigma A')^{-1}\boldsymbol{x}d\boldsymbol{x} = \mathbb{E}[X_A{}'(A\Sigma A')^{-1}X_A] = \mathbb{E}[tr(X_A{}'(A\Sigma A')^{-1}X_A)]$$

$$\mathbb{E}[tr(X_A{}'(A\Sigma A')^{-1}X_A)] = \mathbb{E}[tr((A\Sigma A')^{-1}X_AX_A{}')] = tr[\mathbb{E}(A\Sigma A')^{-1}X_AX_A{}']$$

$$tr[\mathbb{E}(A\Sigma A')^{-1}X_AX_A{}'] = tr[(A\Sigma A')^{-1}\mathbb{E}(X_AX_A{}')] = tr[(A\Sigma A')^{-1}AA']$$

That is,

$$\int f_A(x)\boldsymbol{x}'(A\Sigma A')^{-1}\boldsymbol{x}d\boldsymbol{x} = tr[(A\Sigma A')^{-1}AA']$$

Still we have:

$$\int f_A(\boldsymbol{x})\boldsymbol{x}'(AA')^{-1}\boldsymbol{x}d\boldsymbol{x} = \mathbb{E}[X_A{}'(AA')^{-1}X_A] = \mathbb{E}[tr(X_A{}'(AA')^{-1}X_A)]$$

$$\mathbb{E}[tr(X_A{}'(AA')^{-1}X_A)] = \mathbb{E}[tr((AA')^{-1}X_AX_A{}')] = tr[\mathbb{E}(AA')^{-1}X_AX_A{}']$$

$$tr[\mathbb{E}(AA')^{-1}X_AX_A{}'] = tr[(AA')^{-1}\mathbb{E}X_AX_A{}'] = tr[(AA')^{-1}AA'] = trI_r = r$$

That is,

$$\int f_A(\boldsymbol{x})\boldsymbol{x}'(AA')^{-1}\boldsymbol{x}d\boldsymbol{x} = r$$

Summing the results together, we have:

$$D_{\mathrm{KL}}(f_A\|g_A) = \frac{1}{2}\log\frac{|A\Sigma A'|}{|AA'|} + \frac{1}{2}tr[(A\Sigma A')^{-1}AA'] - \frac{1}{2}r$$

Solution to problem 3 (partial solution)

We define a function $P(x) := \log(x) + \frac{1}{x},\ x > 0$. Suppose the covariance matrix Σ is non-degenerate, then the covariance matrix Σ is positive definite (Generally speaking, Σ is semi-positive definite, because $a'\Sigma a = \mathrm{var}(a'X) \geq 0, \forall a$. However, we assume that Σ is non-degenerate, so it doesn't have eigenvalues of 0, hence Σ is positive definite). Assume $\lambda_1, \dots, \lambda_n > 0$ are n eigenvalues of Σ, and $v_1, \dots, v_n \in \mathbb{R}^{n\times 1}$ are eigenvectors corresponding to the eigenvalues $\lambda_1, \dots, \lambda_n$. Without loss of generality, we assume that each eigenvector is of unit length, that is, $\|v_i\|_2^2 = v_i'v_i = 1,\ 1 \leq i \leq n$, and we also assume that the eigenvalues are in such an order that $P(\lambda_1) \geq P(\lambda_2) \geq \cdots \geq P(\lambda_n)$. Then I assert that the K-L divergence $D_{\mathrm{KL}}(f_A\|g_A)$ reaches its maximum, $\frac{1}{2}\left[\sum_{i=1}^{r}\left(\log(\lambda_i) + \frac{1}{\lambda_i}\right) - r\right] = \frac{1}{2}[\sum_{i=1}^{r}P(\lambda_i) - r]$, when $\mathrm{rank}(A) = r$ and the linear space spanned by A's r row vectors is identical to the linear space spanned by $v_1, \dots v_r$ (the linear space spanned by $v_1, \dots v_r$ might not be unique, provided that $P(\lambda_r) = P(\lambda_{r+1})$).

Since $\sum_{i=1}^{r}P(\lambda_i) - r = \sum_{i=1}^{r}(P(\lambda_i) - 1)$ and $x = 1$ is the only point where $P(x)$ reaches its minimum, $P(x) = 1$, as r increases, the maximum of K-L divergence will continue to increase until all the remaining eigenvalues, $\lambda_{r+1}, \dots, \lambda_n$, satisfy $\lambda_{r+1} = \cdots = \lambda_n = 1$.

I am still working towards a proof of the above-mentioned answer. I am making some progress, but the results are not conclusive. The above-mentioned answer is still my conjecture which needs further study to confirm its correctness.

Here are some of the progress I have made regarding problem 3:

1. if $r = 1$, it is easy to verify that the KL-divergence reaches its maximum when $A \in \mathbb{R}^{1\times n}$

is a row vector satisfying that $A' = kv_1$ $(k \neq 0, \text{and } v_1 \text{ might not be unique if } P(\lambda_1) = P(\lambda_2))$.

First, we obtain the eigenvalue decomposition of the covariance matrix, Σ.

Assume λ_i and λ_j are two different, non-zero eigenvalues of Σ, and $v_i, v_j \in \mathbb{R}^{n\times 1}$ are the corresponding eigenvectors, we have:

$$\Sigma v_i = \lambda_i v_i$$
$$\Sigma v_j = \lambda_j v_j$$

$v_j'\Sigma v_i$ and $v_i'\Sigma v_j$ are scalars, and:

$$v_j'\Sigma v_i = \lambda_i v_j' v_i$$
$$v_i'\Sigma v_j = \lambda_j v_i' v_j$$

As the covariance matrix Σ is symmetric, $\Sigma = \Sigma'$, and $\left(v_j'\Sigma v_i\right)' = v_i'\Sigma v_j$, we have:

$$\left(v_j'\Sigma v_i\right)' = \lambda_i\left(v_j' v_i\right)' = \lambda_i v_i' v_j = \lambda_j v_i' v_j = v_i'\Sigma v_j$$

Since $\lambda_i \neq \lambda_j$, and we have assumed in previous discussion that Σ is positive definite, which means that $\lambda_i \neq 0$ and $\lambda_j \neq 0$, we have:

$$v_i' v_j = 0$$

That is to say, the eigenvector v_i is orthogonal to v_j, if they belong to different eigenvalues $\lambda_i \neq \lambda_j$.

If $\lambda_i = \lambda_j$, in the linear subspace spanned by all eigenvectors of eigenvalue λ_i, we can obtain an orthonormal basis from Gram-Schmidt procedure.

To conclude, if $\Sigma \in \mathbb{R}^{n\times n}$ has k distinct eigenvalues (if $\lambda_i = \lambda_j$ then they are counted as one), then linear space $\mathbb{R}^n$ can be decomposed into the direct sum of k invariant linear subspace, $\mathbb{R}^n = V_1 \oplus \ldots \oplus V_k$. Each invariant linear subspace, V_i, is corresponding to an eigenvalue λ_i of Σ, which means that $\forall v \in V_i$, $1 \leq i \leq k$, we have $\Sigma v = \lambda_i v$. Each invariant linear subspace is invariant under the linear transformation of Σ, which means that $\forall v \in V_i, 1 \leq i \leq k,$ we have $\Sigma v \in V_i$.

From previous discussion we know that different invariant linear subspaces $V_i, V_j (i \neq j)$ are orthogonal to each other, and we can always obtain an orthonormal basis of any invariant linear subspace, $V_i (1 \leq i \leq k)$. As a result, we can obtain an orthonormal basis of $\mathbb{R}^n$, $v_1, \ldots v_n$, satisfying the following conditions: i) each v_i is an eigenvector belonging to the eigenvalue λ_i, ii) different eigenvectors, $v_i, v_j, i \neq j$, are orthogonal to each other, iii) all eigenvectors are of unit length, which means that their ℓ^2-norms equal 1.

Hence, we can find the eigenvectors, $v_1, \ldots v_n \in \mathbb{R}^{n\times 1}$, and they form an orthonormal basis of $\mathbb{R}^n$. So we have $v_j' v_i = \delta_{ij}$ and $v_j'\Sigma v_i = \delta_{ij}\lambda_i$, δ_{ij} is the Kronecker symbol:

$$\delta_{ij} = \begin{cases} 1 & if \;\; i = j \\ 0 & otherwise \end{cases}$$

Since $v_j' v_i = \delta_{ij}$, we have the matrix $V = (v_1 \quad v_2 \quad \ldots \quad v_n)$, $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix:

$$V'V = \begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix} (v_1 \quad v_2 \quad \ldots \quad v_n) = I_n$$

Because $v_j'\Sigma v_i = \delta_{ij}\lambda_i$, we have:

$$V'\Sigma V = \begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix} \Sigma (v_1 \quad v_2 \quad \dots \quad v_n) = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

We denote by Λ the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \dots \lambda_n) = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$.

Making use of the orthogonality of the matrix V, we obtain the eigenvalue decomposition of Σ:

$$\Sigma = (VV')\Sigma(VV') = V(V'\Sigma V)V' = V\Lambda V' = (v_1 \quad v_2 \quad \dots \quad v_n) \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix}$$

Suppose $r = 1$, and $A \in \mathbb{R}^{r \times n}$ is of rank-1. Because $v_1, \dots v_n \in \mathbb{R}^{n \times 1}$ form an orthogonal basis of $\mathbb{R}^n$, we assume that $A \in \mathbb{R}^{1 \times n}$ can be represented by $\mathrm{A} = \varepsilon_1 v_1' + \cdots + \varepsilon_n v_n'$.

Without loss of generality (I will prove this in later part), we assume that $AA' = \varepsilon_1^2 + \cdots + \varepsilon_n^2 = 1$. Then we have $\mathrm{A}\Sigma\mathrm{A}' = \varepsilon_1^2\lambda_1 + \cdots + \varepsilon_n^2\lambda_n \in \mathbb{R}^{1 \times 1}$ is a scalar and

$$D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\left(\log|A\Sigma A'| + \frac{1}{A\Sigma A'} - 1\right) = \frac{1}{2}(P(A\Sigma A') - 1)$$

It is easy to see that the function $P(x)$ is monotonously decreasing in the interval of $(0,1)$ and it is monotonously increasing in the interval of $(1, +\infty)$. Since $\varepsilon_1^2 + \cdots + \varepsilon_n^2 = 1$, we have $\mathrm{A}\Sigma\mathrm{A}' = \varepsilon_1^2\lambda_1 + \cdots + \varepsilon_n^2\lambda_n \in \left[\min_{1 \le i \le n}\{\lambda_i\}, \max_{1 \le i \le n}\{\lambda_i\}\right]$. By the monotonicity of the function $P(x)$, we deduce that $P(A\Sigma A')$ will reach its maximum when $A\Sigma A'$ equals one of the eigenvalues of Σ. To maximize $D_{\mathrm{KL}}(f_A \| g_A)$, we just have to maximize $P(A\Sigma A')$. To maximize $P(A\Sigma A')$, we can simply let $\varepsilon_1 = 1$ and $\varepsilon_2 = \cdots = \varepsilon_n = 0$, since we have assumed that $P(\lambda_1) \ge P(\lambda_2) \ge \cdots \ge P(\lambda_n)$.

Hence, we deduce that $D_{\mathrm{KL}}(f_A \| g_A)$ will reach its maximum in the case of $r = 1$, when $\mathrm{A}' = \alpha v_1$. α is a non-zero scalar, and $v_1 \in \mathbb{R}^{n \times 1}$ is the eigenvector corresponding to the eigenvalue λ_1. A' is identical to v_1 up to a scale factor, and A' will have multiple choices if there exists $k > 1$ and $\mathrm{P}(\lambda_1) = \cdots = \mathrm{P}(\lambda_k) > \mathrm{P}(\lambda_{k+1}) \ge \cdots \ge \mathrm{P}(\lambda_n)$.

2. In the case of $r > 1$, I confirmed that $A \in \mathbb{R}^{r \times n}$ should be of rank r. As the K-L divergence $D_{\mathrm{KL}}(f_A \| g_A)$ remains invariant under certain scalar operations or linear transformations of A, I found out that we can focus only on the cases when $A \in \mathbb{R}^{r \times n}$ is composed of r orthogonal row vectors of unit length.

2.1. First, we can rule out the 'singular' cases when the rank of $A \in \mathbb{R}^{r \times n}$ is less than r.

If $\mathrm{rank}(A) < r$, according to the inequality $\mathrm{rank}(AB) \le \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$, we can deduce that $\mathrm{rank}(AA') \le \mathrm{rank}(A) < r$ and $\mathrm{rank}(A\Sigma A') \le \mathrm{rank}(A) < r$, which means that $AA' \in \mathbb{R}^{r \times r}$ and $A\Sigma A' \in \mathbb{R}^{r \times r}$ are degenerate. So $|AA'| = 0$, $|A\Sigma A'| = 0$ and $A\Sigma A'$ is not invertible. In this 'singular' case, the expression of K-L divergence will become meaningless:

$$D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\log\frac{|A\Sigma A'|}{|AA'|} + \frac{1}{2}tr[(A\Sigma A')^{-1}AA'] - \frac{1}{2}r$$

So we will always disregard this 'singular' cases of $\mathrm{rank}(A) < r$ in the following discussions.

2.2. Second, we can show that the K-L divergence remains invariant under scalar operations of $\tilde{A} = \alpha A, \alpha \neq 0$, which is the reason why we can always assume $|AA'| = 1$ without loss of generality.

$$\begin{aligned} D_{\mathrm{KL}}(f_{\tilde{A}} \| g_{\tilde{A}}) &= \frac{1}{2}\log\frac{\left|\tilde{A}\Sigma\tilde{A}'\right|}{\left|\tilde{A}\tilde{A}'\right|} + \frac{1}{2}tr\left[\left(\tilde{A}\Sigma\tilde{A}'\right)^{-1}\tilde{A}\tilde{A}'\right] - \frac{1}{2}r \\ &= \frac{1}{2}\log\frac{\alpha^2|A\Sigma A'|}{\alpha^2|AA'|} + \frac{1}{2}tr[\alpha^{-2}(A\Sigma A')^{-1}\alpha^2 AA'] - \frac{1}{2}r \end{aligned}$$

It is easy to see that $D_{\mathrm{KL}}(f_{\tilde{A}} \| g_{\tilde{A}}) = D_{\mathrm{KL}}(f_A \| g_A)$ if $\tilde{A} = \alpha A, \alpha \neq 0$. So for any $A \in \mathbb{R}^{r\times n}$ of rank r, $|AA'| \neq 0$, we can always let $\tilde{A} = \frac{1}{\sqrt{|AA'|}}A$, then we have $\left|\tilde{A}\tilde{A}'\right| = 1$. Because $D_{\mathrm{KL}}(f_{\tilde{A}} \| g_{\tilde{A}}) = D_{\mathrm{KL}}(f_A \| g_A)$, by this scalar operation on A we won't miss any cases when K-L divergence reaches its maximum.

2.3. Third, the K-L divergence remains invariant under certain linear transformation of A.

We prove that if we add k times of the j-th row of A to the i-th row of A, then the K-L divergence remains the same.

We denote by $T_k(i,j)$ the above-mentioned linear transformation. It is easy to see that the linear transformation $T_k(i,j)$ on A can be represented as A left-multiplied by a matrix. We still use $T_k(i,j) \in \mathbb{R}^{r\times r}$ to denote this matrix if it doesn't cause confusion in the context.

$$T_k(i,j) = \begin{pmatrix} 1 & & \cdots & & 0 \\ & \ddots & & & \\ & & 1 \cdots k & & \\ \vdots & & \ddots \vdots & & \vdots \\ & & & 1 & \\ & & & & \ddots \\ 0 & & \cdots & & 1 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow\ i-th\ row \\ \\ \leftarrow\ j-th\ row \\ \\ \\ \end{matrix}$$

$$\begin{matrix} \uparrow & \uparrow \\ i-th & j-th \\ column & column \end{matrix}$$

For brevity, we denote $T_k(i,j)$ by T if it doesn't cause confusion in the context.

Let $\tilde{A} = TA$, we have:

$$\begin{aligned} D_{\mathrm{KL}}(f_{\tilde{A}} \| g_{\tilde{A}}) &= \frac{1}{2}\log\frac{|TA\Sigma A'T'|}{|TAA'T'|} + \frac{1}{2}tr[(TA\Sigma A'T')^{-1}TAA'T'] - \frac{1}{2}r \\ &= \frac{1}{2}\log\frac{|T||A\Sigma A'||T'|}{|T||AA'||T'|} + \frac{1}{2}tr[(T')^{-1}(A\Sigma A')^{-1}T^{-1}TAA'T'] - \frac{1}{2}r \\ &= \frac{1}{2}\log\frac{|A\Sigma A'|}{|AA'|} + \frac{1}{2}tr[(T')^{-1}(A\Sigma A')^{-1}AA'T'] - \frac{1}{2}r \\ &= \frac{1}{2}\log\frac{|A\Sigma A'|}{|AA'|} + \frac{1}{2}tr[T'(T')^{-1}(A\Sigma A')^{-1}AA'] - \frac{1}{2}r \\ &= \frac{1}{2}\log\frac{|A\Sigma A'|}{|AA'|} + \frac{1}{2}tr[(A\Sigma A')^{-1}AA'] - \frac{1}{2}r \end{aligned}$$

So $D_{\mathrm{KL}}(f_{\tilde{A}}\|g_{\tilde{A}}) = D_{\mathrm{KL}}(f_A\|g_A)$, which means that K-L divergence remains invariant under certain linear transformations on A.

2.4. From section 2.3. we learned that we can always obtain $\tilde{A} \in \mathbb{R}^{r\times n}$ whose r row vectors are orthogonal to each other, since we can perform the linear transformations on the rows of A without affecting the K-L divergence. Now we show that we can always obtain $\tilde{A} \in \mathbb{R}^{r\times n}$ whose r row vectors are not only orthogonal to each other, but also of unit length.

We define a diagonal matrix $U = \begin{bmatrix} u_1 & & \\ & \ddots & \\ & & u_r \end{bmatrix} \in \mathbb{R}^{r\times r}$ satisfying $u_1 \dots u_r = 1$. Then UA is a linear transformation of A multiplying the i-th row of A by u_i times. We notice that U will not change the determinant of A, so U just serves to rescale each row vector of A without changing $|A|$. We can show that the K-L divergence will not change under these kind of linear transformations on A.

Let $\tilde{A} = UA$, we have $\left|\tilde{A}\Sigma\tilde{A}'\right| = |UA\Sigma A'U'| = |U||A\Sigma A'||U|$. As $u_1 \dots u_r = 1$, we have $|U| = 1$ and $\left|\tilde{A}\Sigma\tilde{A}'\right| = |A\Sigma A|$. Similarly, we have $\left|\tilde{A}\tilde{A}'\right| = |UAA'U'| = |U||AA'||U| = |AA'|$.

For $tr\left[\left(\tilde{A}\Sigma\tilde{A}'\right)^{-1}(\tilde{A}\tilde{A}')\right]$, we have $tr\left[\left(\tilde{A}\Sigma\tilde{A}'\right)^{-1}(\tilde{A}\tilde{A}')\right] = tr[U'^{-1}(A\Sigma A')^{-1}U^{-1}UAA'U'] = tr[U'^{-1}(A\Sigma A')^{-1}AA'U'] = tr[U'U'^{-1}(A\Sigma A')^{-1}AA'] = tr[(A\Sigma A')^{-1}AA']$. Summarizing the above-mentioned results, we have $D_{\mathrm{KL}}(f_{\tilde{A}}\|g_{\tilde{A}}) = D_{\mathrm{KL}}(f_A\|g_A)$ if $\tilde{A} = UA$.

Now we can show that we can focus only on the cases when $A \in \mathbb{R}^{r\times n}$ is composed of r orthonormal row vectors.

First, according to section 2.3., we can perform linear transformations on the rows of A to obtain $\tilde{A}$ whose r row vectors are orthogonal to each other. Then, according to section 2.2., we can rescale the determinant of A to make sure that $\left|\tilde{A}\tilde{A}'\right| = 1$. At last, we can perform the linear transformation mentioned in section 2.4. to rescale each row component of A to make sure that every row vector of $\tilde{A}$ is of unit length. Since all those operations keep the K-L divergence invariant, we can focus only on the cases when $A \in \mathbb{R}^{r\times n}$ is composed of r orthonormal row vectors. Hence, $AA' = I_r$ and $|AA'| = 1$. So we will use the equation $D_{\mathrm{KL}}(f_A\|g_A) = \frac{1}{2}\log|A\Sigma A'| + \frac{1}{2}tr[(A\Sigma A')^{-1}] - \frac{1}{2}r$ to compute K-L divergence without further notice.

Conversely, if the r row vectors of $A \in \mathbb{R}^{r\times n}$ span a linear space that is identical to the linear space spanned by $\tilde{A}$'s r row vectors, then $D_{\mathrm{KL}}(f_{\tilde{A}}\|g_{\tilde{A}})$ and $D_{\mathrm{KL}}(f_A\|g_A)$ will be the same.

3. If $r = n$, we show that $D_{\mathrm{KL}}(f_A\|g_A)$ will reach its maximum of $\frac{1}{2}\left[\sum_{i=1}^{n}\left(\log(\lambda_i) + \frac{1}{\lambda_i}\right) - n\right]$ as long as $A \in \mathbb{R}^{n\times n}$ is of rank n.

In the discussion of section 2, we have ruled out the 'singular' cases when $\mathrm{rank}(A) < r = n$. Now we show that K-L divergence will reach the maximum for every non-degenerate matrix $A \in \mathbb{R}^{r\times n}$ in the case of $r = n$.

According to section 2.2., 2.3., 2.4., we have learned that from a non-degenerate matrix A of

rank n, we can obtain a matrix whose n row vectors can form an orthonormal basis of $\mathbb{R}^n$. In future discussion, we still denote this matrix by A since the K-L divergence stays the same.

We denote by $w_1, \dots w_n \in \mathbb{R}^{n\times 1}$ the above-mentioned orthonormal basis of $\mathbb{R}^n$, so the matrix A can be represented by:

$$A = \begin{pmatrix} w_1' \\ w_2' \\ \vdots \\ w_n' \end{pmatrix}$$

Recalling from section 1, the covariance matrix Σ is:

$$\Sigma = V\Lambda V' = (v_1 \quad v_2 \quad \dots \quad v_n)\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}\begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix}$$

$v_1, \dots v_n \in \mathbb{R}^{n\times 1}$ is also an orthonormal basis of $\mathbb{R}^n$. So we can find an orthogonal matrix C satisfying:

$$A = \begin{pmatrix} w_1' \\ w_2' \\ \vdots \\ w_n' \end{pmatrix} = C\begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix} = CV'$$

So $A\Sigma A' = CV'V\Lambda V'VC' = C\Lambda C'$. Since C is orthogonal, we have $|A\Sigma A'| = |C||\Lambda||C'| = |\Lambda|$. Moreover, $(A\Sigma A')^{-1} = (C\Lambda C')^{-1} = (C')^{-1}\Lambda^{-1}C^{-1} = C\Lambda^{-1}C'$, so $tr[(A\Sigma A')^{-1}] = tr[C\Lambda^{-1}C'] = tr[C'C\Lambda^{-1}] = tr[\Lambda^{-1}]$.

Hence, when $r = n$, we have:

$$D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\log|A\Sigma A'| + \frac{1}{2}tr[(A\Sigma A')^{-1}] - \frac{1}{2}n = \frac{1}{2}\log|\Lambda| + \frac{1}{2}tr[\Lambda^{-1}] - \frac{1}{2}n$$

Since Λ is a diagonal matrix, it is easy to compute:

$$D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\log|\Lambda| + \frac{1}{2}tr[\Lambda^{-1}] - \frac{1}{2}n = \frac{1}{2}\sum_{i=1}^{n}\left(log\lambda_i + \frac{1}{\lambda_i} - 1\right)$$

Remember that the above-mentioned equation is derived only under a mild assumption that $\mathrm{rank}(A) = n$. It means that, in the case of $r = n$, the K-L divergence will always be the maximum of $\frac{1}{2}\sum_{i=1}^{n}\left(log\lambda_i + \frac{1}{\lambda_i} - 1\right)$ as long as A is non-degenerate.

4. Discussion:

From section 1 we deduced that the K-L divergence will reach its maximum of $\frac{1}{2}\left(\log\lambda_1 + \frac{1}{\lambda_1} - 1\right)$ in the case of $r = 1$. From section 3 we derived that the K-L divergence will be the maximum of $\frac{1}{2}\sum_{i=1}^{n}\left(log\lambda_i + \frac{1}{\lambda_i} - 1\right)$ in the case of $r = n$. It is plausible that, for arbitrarily given $r < n$, the K-L divergence will reach the maximum of $\frac{1}{2}\sum_{i=1}^{r}\left(log\lambda_i + \frac{1}{\lambda_i} - 1\right)$ as $P(\lambda_1) \geq P(\lambda_2) \geq \cdots \geq P(\lambda_n)$, $P(x) = \log(x) + \frac{1}{x}$.

From statistical point of view, the eigenvalue decomposition of the covariance matrix Σ:

$$\Sigma = V\Lambda V' = (v_1 \quad v_2 \quad \cdots \quad v_n)\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{bmatrix}\begin{pmatrix}v_1' \\ v_2' \\ \vdots \\ v_n'\end{pmatrix}$$

means that $V'\Sigma V = \Lambda$, that is, $V'\mathbb{E}[(Y-\mathbb{E}Y)(Y-\mathbb{E}Y)']V = \mathbb{E}\left[\left(V'Y-\mathbb{E}(V'Y)\right)\left(V'Y-\mathbb{E}(V'Y)\right)'\right] = \Lambda$. That is to say, according to covariance matrix Σ we can find linear combinations of Y, $V'Y$, so that each linear combination of Y (that is, each component of $V'Y$) is uncorrelated and independent with each other.

As we have seen in the cases of $r = 1$, we should choose $A = \alpha v_1'$ $(\alpha \neq 0)$ to get the maximum of K-L divergence. We should notice that $Y_A = AY = \alpha v_1'Y$, and $v_1'Y$ is the component of $V'Y$ that deviates the most from the covariance matrix of I_n in an information theoretic sense. So, it is plausible, and intuitively appealing, to propose that each component of $V'Y$ contributes independently to K-L divergence. To be specific, the component of $v_i'Y$ contributes an "information gain" of $\log(\lambda_i) + \frac{1}{\lambda_i} - 1$ to the K-L divergence. As the K-L divergence we studied here is concerning the difference between the distribution of $Y \sim \mathcal{N}(0, \Sigma)$ and $X \sim \mathcal{N}(0, I)$, the component of $V'Y$ whose variance is farther from 1 will contribute more to the K-L divergence.

As the dimension *r* grows, the K-L divergence also increases, so we are able to retain more information reflecting the difference between $Y \sim \mathcal{N}(0, \Sigma)$ and $X \sim \mathcal{N}(0, I)$. If we want to compress the information or to represent the information in low rank, then from the point of view of K-L divergence, we should identify the *r* components of $V'Y$ that are corresponding to the *r*-largest items of $P(\lambda_1), \ldots, P(\lambda_r)$.

From section 1 we learned that the choice of A is very limited when $r = 1$. In contrast, according to section 3 virtually all non-degenerate matrix *A* will suffice to lead to K-L divergence's reaching its maximum in the cases of $r = n$. This seems to imply that we should care about the encoding problem when the dimension is low, and we don't have to worry much about the encoding problem if there is "redundancy" in dimensionality.

It is plausible that as *r* increases, we will have a wider range of choices for *A* to maximize the K-L divergence. My conjecture is that the K-L divergence will reach its maximum as long as the linear space spanned by the row vectors of $A \in \mathbb{R}^{r\times n}$ is identical to the linear space spanned by $v_1, \ldots, v_r$. In the next section, we will verify that in the cases of $r < n$, the K-L divergence will reach its hypothetical maximum of $\frac{1}{2}\sum_{i=1}^{r}\left(log\lambda_i + \frac{1}{\lambda_i} - 1\right)$ as long as the above-mentioned requirement for $A \in \mathbb{R}^{r\times n}$ is met.

5. We can verify that, as long as the linear space spanned by the *r* row vectors of $A \in \mathbb{R}^{r\times n}$ is identical to the linear space spanned by $v_1, \ldots, v_r$ which correspond to the *r*-largest items in $P(\lambda_1), P(\lambda_2), \ldots, P(\lambda_n)$, the K-L divergence will reach the value of $\frac{1}{2}\sum_{i=1}^{r}\left(log\lambda_i + \frac{1}{\lambda_i} - 1\right)$, which is supposedly the maximum of K-L divergence when $r < n$.

From section 2 we have learned that we just have to focus on the cases when $A \in \mathbb{R}^{r\times n}$ is composed of *r* orthonormal row vectors. We assume that $w_1, \ldots, w_r \in \mathbb{R}^{n\times 1}$ are *r* orthonormal

vectors and $A = \begin{pmatrix} w_1{}' \\ w_2{}' \\ \vdots \\ w_r{}' \end{pmatrix}$. From $w_1, \dots, w_r \in \mathbb{R}^{n\times 1}$ we can obtain an orthonormal basis of $\mathbb{R}^n$, $w_1, \dots, w_r, w_{r+1}, \dots, w_n$. We denote $\begin{pmatrix} w_{r+1}{}' \\ w_{r+2}{}' \\ \vdots \\ w_n{}' \end{pmatrix}$ by $A_c \in \mathbb{R}^{(n-r)\times n}$, so $\begin{pmatrix} A \\ A_c \end{pmatrix} = \begin{pmatrix} w_1{}' \\ w_2{}' \\ \vdots \\ w_n{}' \end{pmatrix}$. Similar to the discussion in section 3, we can find an orthogonal matrix $C \in \mathbb{R}^{n\times n}$ satisfying $\begin{pmatrix} A \\ A_c \end{pmatrix} = \begin{pmatrix} w_1' \\ w_2' \\ \vdots \\ w_n' \end{pmatrix} = C \begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix} = CV'$.

Sometimes we use block matrices to denote C:

$$C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

where $C_1 \in \mathbb{R}^{r\times n}$, $C_2 \in \mathbb{R}^{(n-r)\times n}$, $C_{11} \in \mathbb{R}^{r\times r}$, $C_{12} \in \mathbb{R}^{r\times(n-r)}$, $C_{21} \in \mathbb{R}^{(n-r)\times r}$, , $C_{22} \in \mathbb{R}^{(n-r)\times(n-r)}$.

We use the notations that are similar to section 3:

$$\Sigma = V\Lambda V' = (v_1 \quad v_2 \quad \dots \quad v_n) \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{pmatrix} v_1' \\ v_2' \\ \vdots \\ v_n' \end{pmatrix}$$

And we use block matrices to denote Λ:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}$$

where $\Lambda_1 = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \end{pmatrix} \in \mathbb{R}^{r\times r}$, $\Lambda_2 = \begin{pmatrix} \lambda_{r+1} & & & \\ & \lambda_{r+2} & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \in \mathbb{R}^{(n-r)\times(n-r)}$.

We have:

$$\begin{pmatrix} A \\ A_c \end{pmatrix} \Sigma (A' \quad A_c{}') = \begin{pmatrix} A\Sigma A' & A\Sigma A_c{}' \\ A_c\Sigma A' & A_c\Sigma A_c{}' \end{pmatrix} = CV'V\Lambda V'VC' = C\Lambda C' = \begin{pmatrix} C_1\Lambda C_1{}' & C_1\Lambda C_2{}' \\ C_2\Lambda C_1{}' & C_2\Lambda C_2{}' \end{pmatrix}$$

We can see that to study $A\Sigma A'$ is to study $C_1\Lambda C_1{}'$.

It is easy to verify that, if $C_{12} = 0$, then $C_{21} = 0$, C_{11}, C_{22} are orthogonal sub-matrices, and the K-L divergence $D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\log|\Lambda_1| + \frac{1}{2}tr[(\Lambda_1)^{-1}] - \frac{1}{2}r$.

If $C_{12} = 0$, then:

$$CC' = I_n = \begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} = \begin{pmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} C_{11}{}' & C_{21}{}' \\ 0 & C_{22}{}' \end{pmatrix} = \begin{pmatrix} C_{11}C_{11}{}' & C_{11}C_{21}{}' \\ C_{21}C_{11}{}' & C_{21}C_{21}' + C_{22}C_{22}{}' \end{pmatrix}.$$

Since $C_{11}C_{11}' = I_r$, by $\mathrm{rank}(C_{11}C_{11}') = r \le \mathrm{rank}(C_{11})$ we have $\mathrm{rank}(C_{11}) = r$. So C_{11} is non-degenerate, the only solution to $C_{11}\boldsymbol{x} = \mathbf{0}$ is $\boldsymbol{x} = \mathbf{0}$. So we can deduce $C_{21}' = \mathbf{0}$ from $C_{11}C_{21}' = \mathbf{0}$. Hence we proved that $C_{21} = \mathbf{0}$. Since $C_{21} = \mathbf{0}$, it is trivial to observe that $C_{11}C_{11}' =$

I_r and $C_{22}C_{22}' = I_{(n-r)\times(n-r)}$. So C_{11}, C_{22} are orthogonal sub-matrices.

Then we have:

$$A\Sigma A' = C_1 \Lambda C_1' = (C_{11} \quad 0)\begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix}\begin{pmatrix}C_{11}'\\ 0\end{pmatrix} = C_{11}\Lambda_1 C_{11}'$$

Hence, the K-L divergence is:

$$\begin{aligned} D_{\mathrm{KL}}(f_A \| g_A) &= \frac{1}{2}\log|A\Sigma A'| + \frac{1}{2}tr[(A\Sigma A')^{-1}] - \frac{1}{2}r \\ &= \frac{1}{2}\log|C_{11}||\Lambda_1||C_{11}'| + \frac{1}{2}tr[C_{11}'^{-1}\Lambda_1^{-1}C_{11}^{-1}] - \frac{1}{2}r \\ &= \frac{1}{2}\log|\Lambda_1| + \frac{1}{2}tr[C_{11}\Lambda_1^{-1}C_{11}'] - \frac{1}{2}r = \frac{1}{2}log|\Lambda_1| + \frac{1}{2}tr(\Lambda_1^{-1}) - \frac{1}{2}r \end{aligned}$$

Since Λ_1^{-1} is a diagonal matrix, it is easy to compute:

$$D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\sum_{i=1}^{r}\log(\lambda_i) + \frac{1}{2}\sum_{i=1}^{r}\frac{1}{\lambda_i} - \frac{1}{2}r$$

When $C_{12} = 0$ and C_{11} is an orthogonal matrix, the r row vectors of $A \in \mathbb{R}^{r\times n}$ are orthonormal vectors that span a linear space which is identical to the linear space spanned by the r eigenvectors of Σ, $v_1, \dots, v_r$, which correspond to the r largest items in $P(\lambda_1), \dots P(\lambda_n)$.

As we have learned in section 2 that linear transformations on the row vectors of $A \in \mathbb{R}^{r\times n}$ will not change the K-L divergence, we deduce that any $A \in \mathbb{R}^{r\times n}$ whose r row vectors span the same linear space as the one spanned by $v_1, \dots, v_r$ will suffice to maximize the K-L divergence, if we can prove that $D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\sum_{i=1}^{r}\log(\lambda_i) + \frac{1}{2}\sum_{i=1}^{r}\frac{1}{\lambda_i} - \frac{1}{2}r$ is indeed the maximum of K-L divergence in the cases of $r < n$.

6. Some thoughts towards a possible proof:

I am still working to find a proof that $D_{\mathrm{KL}}(f_A \| g_A) = \frac{1}{2}\sum_{i=1}^{r}\log(\lambda_i) + \frac{1}{2}\sum_{i=1}^{r}\frac{1}{\lambda_i} - \frac{1}{2}r$ is actually the maximum in the cases of $r < n$. I have been trying with several methods.

(1) On one hand, it is known that we can find the minimum of $f(G) = N\log|G| + trG^{-1}D$ by using Cholesky's decomposition if G and D are positive definite. One the other hand, to maximize the K-L divergence is to maximize $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ under the constraint of $AA' = I_r$. Realizing the similarity between the objective function of $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ and $f(G) = N\log|G| + trG^{-1}D$, I tried to use Cholesky decomposition to treat the problem.

We have assumed that $\Sigma \in \mathbb{R}^{n\times n}$ is positive definite. Suppose $A \in \mathbb{R}^{r\times n}$ and $\mathrm{rank}(A) = r$, then from linear algebra we know that $A\Sigma A'$ is also positive definite. So by Cholesky decomposition we can uniquely find a lower triangular matrix E satisfying $A\Sigma A' = EE'$. The inverse of E, $F = E^{-1}$, is also a lower triangular matrix. So we have:

$$\begin{aligned} &\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}] = \log|EE'| + tr[(EE')^{-1}] = \log|EE'| + tr[(E')^{-1}E^{-1}] \\ &= \log|F^{-1}||F'^{-1}| + trF'F = trF'F - 2\log|F| = \sum_{i\geq j} f_{ij}^2 - 2\sum_{i=1}^{r}\log f_{ii} \end{aligned}$$

Because $F = E^{-1}$ and E, F are lower triangular matrices, we have $\frac{1}{e_{ii}} = f_{ii}$.

Cholesky decomposition provided an effective way for us to treat the expression, $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$. I am curious if I can proceed to obtain some useful results.

(2) In the section 5, we have learned that:

$$\begin{pmatrix} A \\ A_c \end{pmatrix} \Sigma (A' \quad A_c') = \begin{pmatrix} A\Sigma A' & A\Sigma A_c' \\ A_c\Sigma A' & A_c\Sigma A_c' \end{pmatrix} = CV'V\Lambda V'VC' = C\Lambda C' = \begin{pmatrix} C_1\Lambda C_1' & C_1\Lambda C_2' \\ C_2\Lambda C_1' & C_2\Lambda C_2' \end{pmatrix}$$

Since C is an orthogonal matrix, we know that $\log|C\Lambda C'| + tr[(C\Lambda C')^{-1}] = \log|\Lambda| + tr[\Lambda^{-1}] = \sum_{i=1}^{n} \left(log\lambda_i + \frac{1}{\lambda_i} \right)$. A natural thought is that we can use the formulas for block matrices:

$$\begin{vmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{vmatrix} = |B_{11}||B_{22} - B_{21}B_{11}^{-1}B_{12}|$$

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}^{-1} = \begin{pmatrix} B_{11}^{-1}B_{12}B_{22.1}^{-1}B_{21}B_{11}^{-1} + B_{11}^{-1} & -B_{11}^{-1}B_{12}B_{22.1}^{-1} \\ -B_{22.1}^{-1}B_{21}B_{11}^{-1} & B_{22.1}^{-1} \end{pmatrix} (B_{22.1} = B_{22} - B_{21}B_{11}^{-1}B_{12})$$

to decompose $\log|C\Lambda C'| + tr[(C\Lambda C')^{-1}]$. By decomposition of $\log|C\Lambda C'| + tr[(C\Lambda C')^{-1}]$ we will be able to get $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ and a remaining part. To maximize $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ is to minimize the remaining part. It is very likely that the remaining part will take the form of $f(G) = N\log|G| + trG^{-1}D$ whose minimum is known to us. I tried this method, but I got some contradictory results.

Specifically, we can decompose $\log|C\Lambda C'| + tr[(C\Lambda C')^{-1}]$ and obtain:

$$\log|\Lambda| + tr\Lambda^{-1} = \log|A\Sigma A'| + tr[(A\Sigma A')^{-1}] + \log|U| + tr(U^{-1}) + tr[(C_1\Lambda C_1')^{-1}C_1\Lambda C_2'U^{-1}C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}]$$

where $U = C_2\Lambda C_2' - C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}C_1\Lambda C_2'$.

As $\log|\Lambda| + tr\Lambda^{-1}$ is a constant, to maximize $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ we just need to minimize $\log|U| + tr(U^{-1}) + tr[(C_1\Lambda C_1')^{-1}C_1\Lambda C_2'U^{-1}C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}]$. Because $tr(AB) = tr(BA)$ we have $tr[(C_1\Lambda C_1')^{-1}C_1\Lambda C_2'U^{-1}C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}] = tr[U^{-1}C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2']$. Therefore, we need to minimize:

$$\log|U| + tr(U^{-1}) + tr[U^{-1}C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2']$$

which is:

$$\log|U| + tr\left[U^{-1}(I_{(n-r)\times(n-r)} + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2')\right]$$

There is a lemma indicating the following statement: [1]

If G, D are positive definite matrices, then the minimum of

$$f(G) = N\ln|G| + trG^{-1}D$$

exists, and $f(G)$ reaches its minimum when $G = \frac{1}{N}D$.

So, suppose $I_{(n-r)\times(n-r)} + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2'$ is positive definite and $U = C_2\Lambda C_2' - C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}C_1\Lambda C_2'$ is positive definite, then

$$\log|U| + tr\left[U^{-1}(I_{(n-r)\times(n-r)} + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2')\right]$$

will reach its minimum when $U = I + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2'$.

Substituting U by $C_2\Lambda C_2' - C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}C_1\Lambda C_2'$, we have:

$$C_2\Lambda C_2' - C_2\Lambda C_1'(C_1\Lambda C_1')^{-1}C_1\Lambda C_2' = I + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2'$$

Hence

$$\Lambda - \Lambda C_1'(C_1'\Lambda C_1')^{-1}C_1\Lambda = I + \Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda$$

So

$$\Lambda^{-1} - C_1'(C_1'\Lambda C_1')^{-1}C_1 = \Lambda^{-2} + C_1'(C_1\Lambda C_1')^{-2}C_1$$
$$\Lambda^{-1}(I - \Lambda^{-1}) = C_1'[(C_1\Lambda C_1')^{-1}(I + (C_1\Lambda C_1')^{-1})]C_1$$

Hence, from the condition $U = I + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2'$, we obtain an equation which is determined only by C_1 and Λ. If the assumption that $I_{(n-r)\times(n-r)} + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2'$ and U are positive definite is true, then to minimize $\log|\Lambda| + tr\Lambda^{-1} - \log|A\Sigma A'| - tr[(A\Sigma A')^{-1}]$, we need this equation to hold.

We can rewrite the equation in block matrix:

$$\begin{pmatrix}\Lambda_1^{-1} & 0\\ 0 & \Lambda_2^{-1}\end{pmatrix}\begin{pmatrix}I_r - \Lambda_1^{-1} & 0\\ 0 & I_{n-r} - \Lambda_2^{-1}\end{pmatrix}$$
$$= \begin{pmatrix}C_{11}'\\ C_{12}'\end{pmatrix}\left\{\left[\begin{pmatrix}C_{11} & C_{12}\end{pmatrix}\begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix}\begin{pmatrix}C_{11}'\\ C_{12}'\end{pmatrix}\right]^{-1}\left[\begin{pmatrix}I_r & 0\\ 0 & I_{n-r}\end{pmatrix}\right.\right.$$
$$\left.\left. + \left[\begin{pmatrix}C_{11} & C_{12}\end{pmatrix}\begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix}\begin{pmatrix}C_{11}'\\ C_{12}'\end{pmatrix}\right]^{-1}\right]\right\}\begin{pmatrix}C_{11} & C_{12}\end{pmatrix}$$

However, if we take $C_{12} = 0$, the left side of the equation will not equal the right side, which potentially contradicts my conjecture that $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ will reach its maximum when $C_{12} = 0$.

One possible explanation is that the assumption that $I_{(n-r)\times(n-r)} + C_2\Lambda C_1'(C_1\Lambda C_1')^{-2}C_1\Lambda C_2'$ and U are positive definite is always not true. Meanwhile, there is a possibility that my conjecture might be flawed.

(3) Since $\Sigma = V\Lambda V' = \begin{pmatrix}v_1 & v_2 & \cdots & v_n\end{pmatrix}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{bmatrix}\begin{pmatrix}v_1'\\ v_2'\\ \vdots\\ v_n'\end{pmatrix}$, if we choose $A = \begin{pmatrix}v_1{}'\\ v_2{}'\\ \vdots\\ v_r{}'\end{pmatrix}$, we will have $A\Sigma A' = \Lambda_1$ and $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}] = \sum_{i=1}^{r}\log(\lambda_i) + \sum_{i=1}^{r}\frac{1}{\lambda_i}$. I can prove that this is a Lagrangian stationary point for the optimization problem:

maximize: $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$

subject to: $AA' = I_r$

Being a Lagrangian stationary point is a necessary condition for being an extremal point. However, it is not sufficient to indicate that this is the maximum of K-L divergence.

If I could identify all the Lagrangian stationary points under the constraint of $AA' = I_r$, then I might be able to verify that $A = \begin{pmatrix}v_1{}'\\ v_2{}'\\ \vdots\\ v_r{}'\end{pmatrix}$ is the extremal point. However, at this time, I can only compute the gradients of $\log|A\Sigma A'|$ and $tr[(A\Sigma A')^{-1}]$ under the condition of $A\Sigma A' = \Lambda_1$, and I will have difficulties in computation if I want to compute the gradients when the condition of $A\Sigma A' = \Lambda_1$ is unmet.

Next, we will prove that, when $A = A_0 \in \mathbb{R}^{r\times n}$ satisfies $A_0\Sigma A_0' = \Lambda_1$, then $A = A_0$ is a Lagrangian stationary point.

First, we compute the derivatives of $\log|A\Sigma A'|$ when $A = A_0$.

We denote by $E_{ij} \in \mathbb{R}^{r\times n}$ a matrix whose element in row-*i*, column-*j* is 1 and all the other elements are 0s. We have:

$$\left(A_0 + tE_{ij}\right)\Sigma\left(\mathrm{A}_0 + tE_{ij}\right)' = A_0\Sigma A_0' + tA_0\Sigma E_{ij}' + tE_{ij}\Sigma A_0' + t^2E_{ij}\Sigma E_{ij}'$$

$$= \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \end{pmatrix}$$

$$+ \begin{pmatrix} & & t\lambda_1 a_{1j} & & \\ & & \vdots & & \\ & & t\lambda_{i-1}a_{(i-1)j} & & \\ t\lambda_1 a_{1j} & \dots & 2t\lambda_i a_{ij} + t^2\sigma_{ij} & \cdots & t\lambda_r a_{rj} \\ & & t\lambda_{i+1}a_{(i+1)j} & & \\ & & \vdots & & \\ & & t\lambda_r a_{rj} & & \end{pmatrix} \leftarrow i-th\ row$$

$$\uparrow$$

$$i-th\ column$$

Its determinant is:

$$\left|\left(A_0 + tE_{ij}\right)\Sigma\left(\mathrm{A}_0 + tE_{ij}\right)'\right| = \left(\prod_{i=1}^{r}\lambda_i\right)\left(1 + 2ta_{ij} + \frac{t^2}{\lambda_i}(\sigma_{ij} - \sum_{k\neq i}\lambda_k a_{kj}^2)\right)$$

So we can compute the derivatives of $\log|A\Sigma A'|$ when $A = A_0$:

$$\frac{\partial\log|A\Sigma A'|}{\partial a_{ij}}_{|A=A_0} = \lim_{t\to 0}\frac{\log\left|\left(A_0 + tE_{ij}\right)\Sigma\left(\mathrm{A}_0 + tE_{ij}\right)'\right| - \log|\Lambda_1|}{t}$$

$$= \lim_{t\to 0}\frac{1}{t}\log\left(1 + 2ta_{ij} + \mathcal{O}(t^2)\right) = 2a_{ij}$$

Reorganize the derivatives in the form of a matrix, we have:

$$\nabla_A\log|A\Sigma A'|_{|A=A_0} = 2A$$

Then we compute the derivatives of $tr[(A\Sigma A')^{-1}]$ when $A = A_0$.

Here the problem is how to compute $\left[\left(A_0 + tE_{ij}\right)\Sigma\left(A_0 + tE_{ij}\right)'\right]^{-1}$. Luckily we have a corollary in functional analysis indicating the following:

Suppose *X* is a Banach space, and $T \in \mathcal{B}(X)$ is a bounded linear operator who has a bounded inverse operator, then for any $\Delta T \in \mathcal{B}(X)$ satisfying $\|\Delta T\| < \frac{1}{\|T^{-1}\|}$, the operator $S = T + \Delta T$ has a bounded inverse operator, and we have:

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

Obviously, $\mathbb{R}^r$ is Banach space and matrices in $\mathbb{R}^{r\times r}$ are bounded linear operators. In our case, $T = A_0\Sigma A_0' = \Lambda_1$, and $\Delta T = t\Lambda_1 A_0 E_{ij}' + tE_{ij}A_0'\Lambda_1 + t^2E_{ij}\Sigma E_{ij}'$. As long as *t* is sufficiently small, we have $\|\Delta T\| < \frac{1}{\|T^{-1}\|}$. Hence we can obtain $\left[\left(A_0 + tE_{ij}\right)\Sigma\left(A_0 + tE_{ij}\right)'\right]^{-1}$:

$$\left[\left(A_0 + tE_{ij}\right)\Sigma\left(A_0 + tE_{ij}\right)'\right]^{-1} = \Lambda_1^{-1} - \Lambda_1^{-1}(\Lambda_1^{-1}\Delta T) + \Lambda_1^{-1}(\Lambda_1^{-1}\Delta T)^{-2} - \cdots$$

As before, in the computation of the derivatives, we can disregard all the terms of an order higher than $\mathcal{O}(t)$, which simplifies the computation:

$$tr\left[\left(A_0 + tE_{ij}\right)\Sigma\left(A_0 + tE_{ij}\right)'\right]^{-1} = tr\Lambda_1^{-1} - tr[\Lambda_1^{-1}(\Lambda_1^{-1}\Delta T)] + \mathcal{O}(t^2)$$

So:

$$\frac{\partial tr[(A\Sigma A')^{-1}]}{\partial a_{ij}}_{|A=A_0} = -\lim_{t\to 0}\frac{tr[\Lambda_1^{-1}(\Lambda_1^{-1}\Delta T)]}{t} = -tr\left(AE_{ij}'\Lambda^{-1}\right) - tr(\Lambda_1^{-1}E_{ij}A')$$

Since $\forall A \in \mathbb{R}^{r\times r}$, $tr(\mathrm{A}) = tr(\mathrm{A}')$, we have:

$$\frac{\partial tr[(A\Sigma A')^{-1}]}{\partial a_{ij}}_{|A=A_0} = -2tr\left(AE_{ij}'\Lambda^{-1}\right) = -\frac{2a_{ij}}{\lambda_i}$$

Reorganize the derivatives in the form of a matrix, we have:

$$\nabla_A tr[(A\Sigma A')^{-1}]_{|A=A_0} = -2\Lambda_1^{-1}A$$

Now we consider the Lagrange multipliers. Since the constraint is $AA' = I_r$, the Lagrangian function, $F(A,\mu)$, is:

$$F(A,\mu) = \log|A\Sigma A'| + tr[(A\Sigma A')^{-1}] + G(A,\mu)$$

$$= \log|A\Sigma A'| + tr[(A\Sigma A')^{-1}] + \sum_{i=1}^{r}\mu_{i,i}\left(\sum_{k=1}^{n}a_{ik}^2 - 1\right) + \sum_{i\neq j}\mu_{i,j}\sum_{k=1}^{n}a_{ik}a_{jk}$$

As we have used λ_is to denote the eigenvectors of Σ, in order to avoid confusion we use $\mu_{i,j}$ to denote the Lagrange multipliers.

There are $r\times r$ Lagrange multipliers, $\mu_{i,j}$ $(1\le i\le r, 1\le j\le r)$. For convenience, we can write them in a matrix form M, $\mathrm{M} = \begin{pmatrix}\mu_{1,1} & \cdots & \mu_{1,r}\\ \vdots & \ddots & \vdots \\ \mu_{r,1} & \cdots & \mu_{r,r}\end{pmatrix} \in \mathbb{R}^{r\times r}$. We can also assume that $\mu_{i,j} = \mu_{j,i}$.

Since we have obtained the derivatives of $\log|A\Sigma A'|$ and $tr[(A\Sigma A')^{-1}]$, now we compute the derivatives of the remaining part of $F(A,\mu)$.

$$\frac{\partial}{\partial a_{ij}}G(A,\mu) = \frac{\partial}{\partial a_{ij}}\left[\sum_{k=1}^{r}\mu_{k,k}\left(\sum_{m=1}^{n}a_{km}^2 - 1\right) + \sum_{k\neq l}\mu_{k,l}\sum_{m=1}^{n}a_{km}a_{lm}\right]$$

$$= \frac{\partial}{\partial a_{ij}}\left[\mu_{i,i}a_{ij}^2\right] + \frac{\partial}{\partial a_{ij}}\left[\sum_{l\neq i}\mu_{i,l}a_{ij}a_{lj}\right] + \frac{\partial}{\partial a_{ij}}\left[\sum_{k\neq i}\mu_{k,i}a_{kj}a_{ij}\right]$$

$$= 2\mu_{i,i}a_{ij} + \sum_{l\neq i}\mu_{i,l}a_{lj} + \sum_{k\neq i}\mu_{k,i}a_{kj}$$

Since we have assumed that $\mu_{i,j} = \mu_{j,i}$, we have:

$$\frac{\partial}{\partial a_{ij}}G(A,\mu) = 2\sum_{k=1}^{r}\mu_{i,k}\,a_{kj}$$

Therefore, we can write the gradient of $G(A,\mu)$ in the form of a matrix:

$$\nabla_A G(A,\mu) = 2\mathrm{M}A$$

So, when $A = A_0$, the gradient of $F(A,\mu)$ is:

$$\nabla_A F(A,\mu)_{|A=A_0} = 2(I_r - \Lambda_1^{-1} + \mathrm{M})A$$

Let $\mathrm{M} = \begin{pmatrix} \mu_{1,1} & \cdots & \mu_{1,r} \\ \vdots & \ddots & \vdots \\ \mu_{r,1} & \cdots & \mu_{r,r} \end{pmatrix} = \begin{pmatrix} \frac{1}{\lambda_1} - 1 & & & \\ & \frac{1}{\lambda_2} - 1 & & \\ & & \ddots & \\ & & & \frac{1}{\lambda_r} - 1 \end{pmatrix} = \Lambda_1^{-1} - I_r$, then we have $\nabla_A F(A,\mu)_{|A=A_0} = 0$. That is to say, $A = A_0$, $\mathrm{M} = \Lambda_1^{-1} - I_r$ is a Lagrangian stationary point.

$\log|\cdot|$ is a concave function, but $tr(\cdot^{-1})$ is a convex function, so it is not very likely that $\log|A\Sigma A'| + tr[(A\Sigma A')^{-1}]$ should be either a concave function or a convex function. Besides, I am not sure if I could transform the constraint $AA' = I_r$ to the standard form in convex optimization.

At this time, I am still trying other approaches in order to obtain a solid proof of my conjecture.

Reference:

[1] An Introduction to Multivariate Statistical Analysis (Third Edition), T. W. Anderson, Lemma 3.2.2.