Approximately Maximizing Total Variation 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.

Since $h_{X_A}(x), h_{Y_A}(x)$ are Lesbegue integrable, the total variation can be expressed as the following:

$$D_{TV}\left(h_{X_A}, h_{Y_A}\right) = \frac{1}{2}\int |h_{X_A}(x) - h_{Y_A}(x)| dx$$

$\int |h_{X_A}(x) - h_{Y_A}(x)| dx$ is also the ℓ^1-norm of $|h_{X_A}(x) - h_{Y_A}(x)|$.

As total variation has no closed form for Gaussian distributions. we alternatively seek to maximize its theoretical boundaries. According to [1], we have:

$$\frac{1}{100} \le \frac{D_{TV}(h_{X_A}, h_{Y_A})}{\min\{1, \sqrt{\sum_{i=1}^r \beta_i^2}\}} \le \frac{3}{2}$$

in which $\beta_1, \beta_2, \ldots, \beta_r$ are eigenvalues of $(AA^T)^{-1}(A\Sigma A^T) - I_r$.

Since $D_{TV}(h_{X_A}, h_{Y_A})$ is upper bounded by $\frac{3}{2}\min\{1, \sqrt{\sum_{i=1}^r \beta_i^2}\}$ and lower bounded by $\frac{1}{100}\min\{1, \sqrt{\sum_{i=1}^r \beta_i^2}\}$. Our problem is to find A to maximize:

$$\min\left\{1, \sqrt{\sum_{i=1}^r \beta_i^2}\right\}$$

Solution:

1. Formulation of the Problem:

To maximize $\min\left\{1, \sqrt{\sum_{i=1}^r \beta_i^2}\right\}$, it is sufficient to maximize $\sum_{i=1}^r \beta_i^2$.

Since $\beta_1, \beta_2, \ldots, \beta_r$ are eigenvalues of $(AA^T)^{-1}(A\Sigma A^T) - I_r$, we have:

$$\sum_{i=1}^r \beta_i^2 = tr[((AA^T)^{-1}(A\Sigma A^T) - I_r)^2]$$

So our aim is to maximize $tr[((AA^T)^{-1}(A\Sigma A^T) - I_r)^2]$.

2. Linear invariant of the objective function:

For any invertible matrix $M \in \mathbb{R}^{\mathrm{r}\times\mathrm{r}}$, $\tilde{A} = MA$, we have:

$$tr\left[\left(\left(\tilde{A}\tilde{A}^T\right)^{-1}\left(\tilde{A}\Sigma\tilde{A}^T\right) - I_r\right)^2\right] = tr[((M^T)^{-1}(AA^T)^{-1}M^{-1}MA\Sigma A^T M^T - I_r)^2]$$

$$= tr[((M^T)^{-1}((AA^T)^{-1}A\Sigma A^T - I_r)M^T)^2]$$

$$= tr[(M^T)^{-1}((AA^T)^{-1}A\Sigma A^T - I_r)^2 M^T] = tr[((AA^T)^{-1}(A\Sigma A^T) - I_r)^2]$$

So our objective function is invariant under invertible linear transformations, which means that we can always assume that $A \in \mathbb{R}^{r\times n}$ is composed of orthonormal row vectors. That is, we can always assume $AA^T = I_r$ without loss of generality.

3. Problem Reduction:

With the constraint $AA^T = I_r$, the objective function can be simplified and our problem can be reformulated as:

$$\max \ tr[((A\Sigma A^T) - I_r)^2]$$
$$\text{s.t. } AA^T = I_r$$

Suppose $A\Sigma A^T$'s eigenvalues are $\gamma_1, \dots, \gamma_r, \gamma_1 \geq \cdots \geq \gamma_r$, then our aim is to maximize $\sum_{i=1}^{r}(\gamma_i - 1)^2$.

We can also simplify the problem without imposing the constraint of $AA^T = I_r$.

Without the constraint $AA^T = I_r$, by SVD decomposition we can assume that:

$$A^T = P\begin{pmatrix} M \\ 0 \end{pmatrix}Q = (P_1, P_2)\begin{pmatrix} M \\ 0 \end{pmatrix}Q$$

in which $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{r\times r}$ are orthogonal matrices, $P_1 \in \mathbb{R}^{n\times r}$ and $P_2 \in \mathbb{R}^{n\times(n-r)}$ are submatrices of P, $M \in \mathbb{R}^{r\times r}$ is a diagonal matrix, and $0 \in \mathbb{R}^{(n-r)\times r}$.

According to [1], $(AA^T)^{-1}A\Sigma A^T$ have the same eigenvalues as $(AA^T)^{-\frac{1}{2}}(A\Sigma A^T)(AA^T)^{-\frac{1}{2}}$, so $(AA^T)^{-1}A\Sigma A^T - I_r$ have the same eigenvalues as $(AA^T)^{-\frac{1}{2}}(A\Sigma A^T)(AA^T)^{-\frac{1}{2}} - I_r$.

we have:

$$AA^\mathrm{T} = Q^T(\mathrm{M} \quad 0)\begin{pmatrix} P_1^T \\ P_2^T \end{pmatrix}(P_1, P_2)\begin{pmatrix} M \\ 0 \end{pmatrix}Q = Q^T\mathrm{M}^2Q$$

So $(AA^T)^{-\frac{1}{2}} = Q^T\mathrm{M}^{-1}Q$, $Z = A\Sigma A^T = Q^T(\mathrm{M} \quad 0)P^T\Sigma P\begin{pmatrix} \mathrm{M} \\ 0 \end{pmatrix}Q$. Therefore, we have:

$$(AA^T)^{-\frac{1}{2}}Z(AA^T)^{-\frac{1}{2}} = Q^TP_1^T\Sigma P_1Q$$

So, to study $(AA^T)^{-1}Z - I_r$'s eigenvalues is to study $(AA^T)^{-\frac{1}{2}}Z(AA^T)^{-\frac{1}{2}} - I_r$'s eigenvalues, which is to study $Q^TP_1^T\Sigma P_1Q - I_r$'s eigenvalues. Since Q is orthogonal, $Q^TP_1^T\Sigma P_1Q - I_r$'s eigenvalues are the same as $P_1^T\Sigma P_1 - I_r$'s eigenvalues.

So, to maximize $tr[((AA^T)^{-1}(A\Sigma A^T) - I_r)^2]$ is to maximize $tr[(P_1^T\Sigma P_1 - I_r)^2]$. Because $P \in \mathbb{R}^{n\times n}$ is orthogonal, we have $P_1^TP_1 = I_r$.

Hence, our problem can be reformulated as:

$$\max \ tr[(P_1^T\Sigma P_1 - I_r)^2]$$
$$\text{s.t. } P_1^\mathrm{T}P_1 = I_r$$

Suppose $P_1^TP_1$'s eigenvalues are $\gamma_1, \dots, \gamma_r, \gamma_1 \geq \cdots \geq \gamma_r$, then our aim is to maximize $\sum_{i=1}^{r}(\gamma_i - 1)^2$.

Obviously, the two versions of problem reduction are basically the same.

4. Cauchy Interlacing Theorem:

No matter we use which version of problem reduction, the core problem is to establish the connection between the eigenvalues of Σ and the eigenvalues of $A\Sigma A^T$ (or $P_1^T\Sigma P_1$). This is done by applying Cauchy's interlacing theorem [2][3][4].

Suppose $A\Sigma A^T$'s (or $P_1^T\Sigma P_1$'s) eigenvalues are $\gamma_1, \dots, \gamma_r, \gamma_1 \geq \cdots \geq \gamma_r$ and Σ's eigenvalues are $\lambda_1, \dots, \lambda_n, \lambda_1 \geq \cdots \geq \lambda_n$, then according to Cauchy's interlacing theorem, we have:

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

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 $\sum_{i=1}^{r} (\gamma_i - 1)^2$ and $tr[((AA^T)^{-1}(A\Sigma A^T) - I_r)^2]$.

5. Results:

First, we make some observations.

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

b) $(\gamma_i - 1)^2$ is monotonously decreasing in the interval of $(0,1]$ and monotonously increasing in the interval of $[1, \infty)$. Besides, $(\gamma_i - 1)^2$ is convex. Either by its monotonicity or by its convexity we can prove that $(\gamma_i - 1)^2$ 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* or P_1 by the corresponding eigenvectors, for each $1 \leq i \leq r$ 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 $tr[((AA^T)^{-1}(A\Sigma A^T) - I_r)^2]$, we need to let each $\gamma_i = \lambda_i$ or $\gamma_i = \lambda_{n-r+i}$. That is to say, in order to maximize total variation, we need to choose *r* elements from $\lambda_i (1 \leq i \leq n)$ and assign them to $\gamma_i (1 \leq i \leq r)$.

So, in order to maximize total variation, we need to select the *r* elements from $\lambda_i (1 \leq i \leq n)$ who have the largest value of $|\gamma_i - 1|$. This is the short version of the results.

The long version of the results are as follows:

Case 1: Suppose $\lambda_1 \geq \cdots \geq \lambda_n \geq 1$, then $\sum_{i=1}^{r} (\gamma_i - 1)^2$ yields its maximum by taking $\gamma_1 = \lambda_1$, $\gamma_2 = \lambda_2, \dots, \gamma_r = \lambda_r$.

Case 2: Suppose $1 \geq \lambda_1 \geq \cdots \geq \lambda_n$, then $\sum_{i=1}^{r} (\gamma_i - 1)^2$ yields its maximum by taking $\gamma_1 = \lambda_{n-r+1}$, $\gamma_2 = \lambda_{n-r+2}, \dots, \gamma_r = \lambda_n$.

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 $(\gamma_i - 1)^2$ 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 $|\lambda_i - 1|$ and $|\lambda_{n-r+i} - 1|$. If $|\lambda_i - 1| > |\lambda_{n-r+i} - 1|$, then we let $\gamma_i = \lambda_i$, otherwise we let $\gamma_i = \lambda_{n-r+i}$. This will suffice to approximately obtain the maxima of total variation.

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

6. Construction of A:

From the discussion in the section 5 we have learned that we can construct *A* (or P_1) by the corresponding eigenvectors.

Specifically, in Case 1, we construct A or P_1 by $\begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{pmatrix}$, in Case 2, we construct A or P_1 by $\begin{pmatrix} v_{n-r+1}^T \\ v_{n-r+2}^T \\ \vdots \\ v_n^T \end{pmatrix}$, 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 or P_1 by $\begin{pmatrix} v_1^T \\ \vdots \\ v_i^T \\ v_{n-r+i+1}^T \\ \vdots \\ v_n^T \end{pmatrix}$. This will suffice to approximately obtain the maxima of total variation.

7. Matrix A in general forms:

In section 2 we have learned that our objective function is invariant under invertible linear transformations in $\mathbb{R}^r$. From this we conclude that total variation 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 $|\lambda_1 - 1|, |\lambda_2 - 1|, \dots, |\lambda_n - 1|$.

Reference:

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

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

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

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