Some Observations Regarding the Maximization of Total Variation

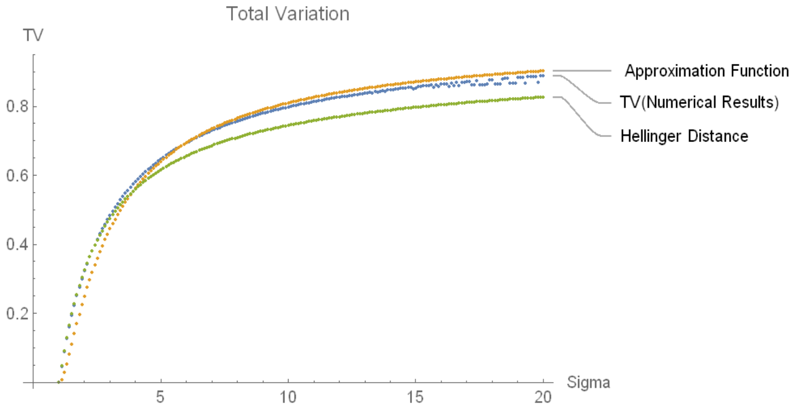
Sihui Wang

1. Univariate Cases:

Because $TV(\mathcal{N}(0,1),\mathcal{N}(0,\sigma^2)) = TV(\mathcal{N}(0,1),\mathcal{N}(0,\frac{1}{\sigma^2}))$, when $r = 1$, to maximize total variation, we should choose the eigenvalue of Σ according to the evaluation function of $\lambda + \frac{1}{\lambda}$.

In univariate cases, Hellinger distance is a good approximation of TV, especially when σ is small. Besides, we can also find an approximation function for $TV(\mathcal{N}(0,1),\mathcal{N}(0,\sigma^2))$:

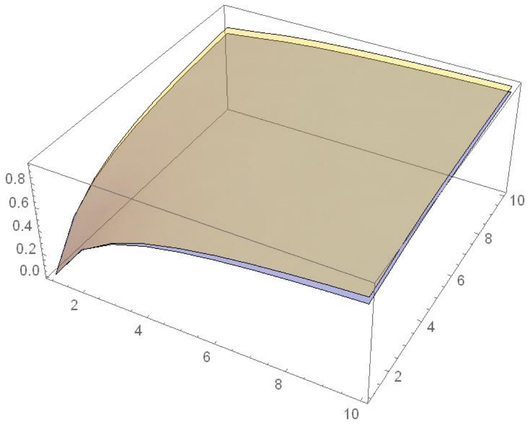
$$TV(\mathcal{N}(0,1),\mathcal{N}(0,\sigma^2)) \approx \begin{cases} \left(1-\frac{1}{\sigma}\right)^2 & \sigma > 1 \\ (1-\sigma)^2 & \sigma < 1 \end{cases}$$



In my numerical computation, this approximation function performs the worst when $\sigma = 1.7$. At that point, the actual integral value is about 0.223365, while the value predicted by the approximation function is 0.140625. But as σ grows, the error of prediction narrows down. For those larger σ, this approximation function can give an accurate prediction of total variation for univariate cases.

2. Two Dimensional Cases:

In two dimensional cases, when $\sigma_1^2 \geq 1, \sigma_2^2 \geq 1$, Hellinger distance is also a good approximation of TV (the orange one above is TV, the blue one below is Hellinger distance):



However, Hellinger distance and TV are different in some ways. For example, if we replace an eigenvalue of Σ, σ_2^2, by $\frac{1}{\sigma_2^2}$, then Hellinger distance is invariant. However, in multivariate cases, this is not true for total variation.

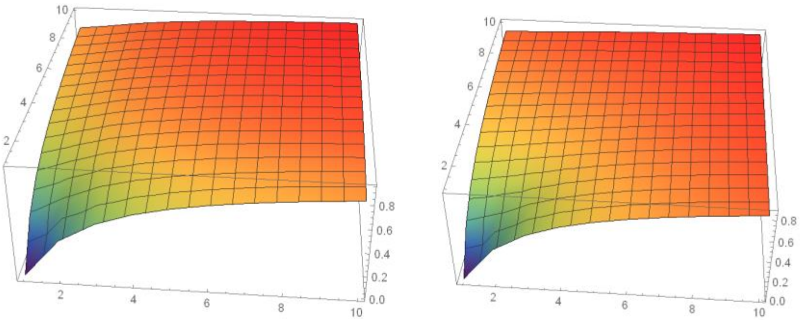
Previously, we have confirmed that, in univariate cases, $\mathcal{N}(0,\sigma^2)$ and $\mathcal{N}\left(0,\frac{1}{\sigma^2}\right)$ are isometric to $\mathcal{N}(0,1)$ under the metric of TV. However, in multivariate cases, this is no longer true. For example, in two dimensional cases, we still have:

$\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{pmatrix}\right)$ and $\mathcal{N}\left(0,\begin{pmatrix}\frac{1}{\sigma_1^2} & \\ & \frac{1}{\sigma_2^2}\end{pmatrix}\right)$ are isometric to $\mathcal{N}(0,I_2)$ under the metric of TV.

But generally speaking, $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{pmatrix}\right)$ and $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \frac{1}{\sigma_2^2}\end{pmatrix}\right)$ are NOT isometric to $\mathcal{N}(0,I_2)$ under the metric of TV.

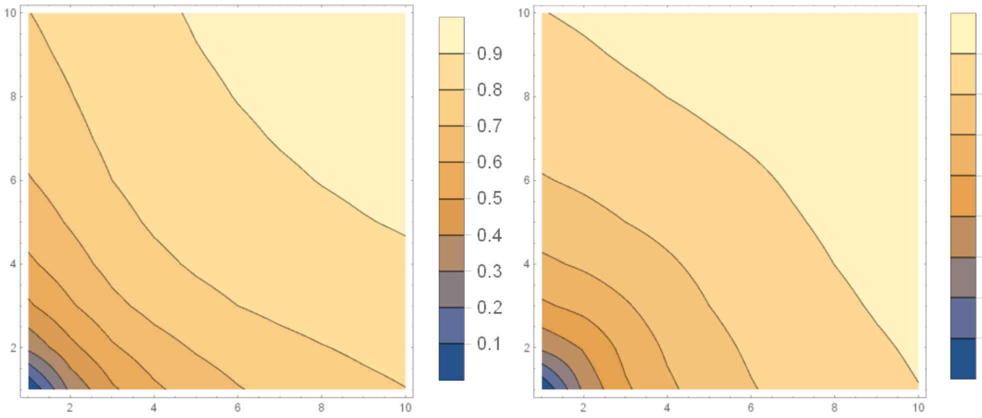
This implies that, although we can decompose the problem of maximization into sub-problems in each dimension in the cases of KL-divergence and Hellinger distance, we CANNOT do the same for TV.

This is total variations in two dimensional cases:

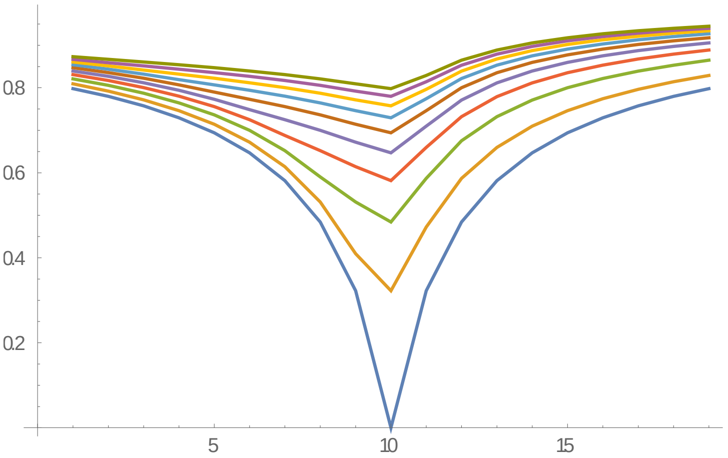


In the left image, σ_1 ranges from 1 to 10 and σ_2 ranges from 1 to 10, whereas in the right image, σ_1 ranges from 1 to 10 and σ_2 ranges from 1 to $\frac{1}{10}$.

These two images seem alike at the first glance, but there are subtle differences.



If $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{pmatrix}\right)$ and $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \frac{1}{\sigma_2^2}\end{pmatrix}\right)$ are isometric to $\mathcal{N}(0,I_2)$ under the metric of TV, then the above contours should be identical. However, the differences in the contours clearly indicate that $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{pmatrix}\right)$ and $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \frac{1}{\sigma_2^2}\end{pmatrix}\right)$ are NOT isometric to $\mathcal{N}(0,I_2)$ under the metric of TV.



In the above image, in each curve from the bottom to the top, σ_1 is chosen as $\sigma_1 = 1, \dots, 10$. For the *x*-axis, σ_2 is chosen as $\sigma_2 = \frac{1}{10}, \dots, \frac{1}{2}, 1, 2, \dots, 10$. The blue curve is symmetric, which indicates that when $\sigma_1 = 1$, σ_2 and $\frac{1}{\sigma_2}$ are equally favored. Other curves are not symmetric, which means that if we have chosen $\sigma_1 > 1$, then in later selections we should favor the larger one, $\sigma_2 (\sigma_2 > 1)$, over the smaller ones, $\frac{1}{\sigma_2} (\sigma_2 > 1)$. Because $\mathcal{N}\left(0,\begin{pmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{pmatrix}\right)$ and $\mathcal{N}\left(0,\begin{pmatrix}\frac{1}{\sigma_1^2} & \\ & \frac{1}{\sigma_2^2}\end{pmatrix}\right)$ are isometric to $\mathcal{N}(0,I_2)$, if we have chosen $\sigma_1 < 1$, then in later selections we should favor the smaller ones, $\frac{1}{\sigma_2} (\sigma_2 > 1)$, over the larger ones, $\sigma_2 (\sigma_2 > 1)$.

3. A Counter Example and Possible Trends in Multivariate Cases:

From the above discussion, we learned that in some cases, Hellinger distance and total variations are quite close. However, in some cases they are different and cannot be maximized simultaneously following the same strategy.

Specifically, if all eigenvalues of Σ are larger than 1, or all eigenvalues of Σ are smaller than 1, then the maximization of total variation is basically the same as the maximization of other f-divergences. But if there are both eigenvalues that are larger than 1 and eigenvalues that are smaller than 1, then the maximization of total variation is not the same as the maximization of Hellinger distance or other f-divergences. The greedy algorithm, which iteratively select an eigenvalue from the remaining ones that contribute the most to f-divergence according to certain evaluation functions, works well for other f-divergences, but it is not a particularly good strategy for maximization of total variation when there are both eigenvalues that are larger than 1 and eigenvalues that are smaller than 1.

In fact, we can find *counterexamples* in two dimensional cases.

Suppose that total dimension $n = 4$, $\Sigma \in \mathbb{R}^{\mathrm{n\times n}}$ has 4 eigenvalues, $\frac{1}{64}, \frac{1}{36}, 49, 49$, and $r = 2$. Following the evaluation function for Hellinger distance and symmetric KL-divergence, $\lambda + \frac{1}{\lambda}$, we should choose $\sigma_1^2 = \frac{1}{64}$ and $\sigma_2^2 = 49$. However, this choice will result in a suboptimal solution for the maximization of total variation. Under this choice, the total variation is: 0.832.

From the above observations we have learned that there is a trend that both of the chosen eigenvalues should be on the same side of 1. So shall we choose $\sigma_1^2 = \frac{1}{64}$ and $\sigma_2^2 = \frac{1}{36}$? This is a good solution, because the total variation under this choice is: 0.902.

However, $\sigma_1^2 = \frac{1}{64}, \sigma_2^2 = \frac{1}{36}$ is only a near optimal solution. Eventually we find the total variation will reach the maxima of 0.903 at $\sigma_1^2 = 49, \sigma_2^2 = 49$.

In one dimensional cases, we should definitely choose $\sigma_1^2 = \frac{1}{64}$. But in two dimensional cases, we can no longer choose the eigenvalues according to the evaluation function, $\lambda + \frac{1}{\lambda}$. Previously, we have observed that if we have chosen one smaller eigenvalue, then in later selection process we should favor the smaller ones over the larger ones. However, this cannot guarantee that we obtain the optimal solution. As we have seen, $\sigma_1^2 = \frac{1}{64}, \sigma_2^2 = \frac{1}{36}$ is only near optimal. The real optimal solution, on the other side, is the one we won't even consider it at the very beginning according to greedy algorithms. This means that the optimal solution to the maximization of total variation might be delicate and diverse. They might depend on the specific configuration of the eigenvalues, and there is no guarantee that greedy algorithms will obtain an optimal solution.

Possible Trends For Total Variation Since we only make observations in numerical tests when $r = 1$ and $r = 2$, we cannot conclude any general principles for total variations in high dimensional cases. However, we can observe some trends that are possibly true for higher dimensional cases.

1) Compared with the cases when all the selected eigenvalues are on the same side of 1, the cases when eigenvalues are distributed on both side of 1 tend to be suboptimal.

2) Observing the contour map on the left, we find that the contours tend to be convex when selected eigenvalues are large. So, to maximize TV, two medium-to-large eigenvalues are sometimes better than one large and one medium eigenvalues.

4. Total Variation in High Dimension:

1) if Σ is only slightly different from I_n in each dimension:

We have confirmed that, if Σ is slightly different from I_n in each dimension, then Hellinger distance tends to be close to 1 when the dimension is very high. As total variation is under bounded by $H^2(P||Q)$, we conclude that total variation also tends to be close to 1 if Σ is only slightly different from I_n in each dimension.

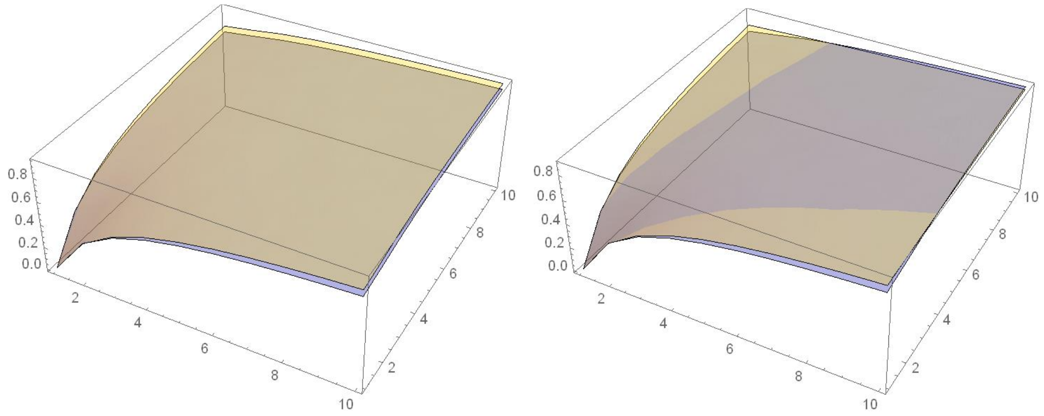
2) if Σ's eigenvalues are uniformly distributed in an interval $\left[\frac{1}{a}, \mathrm{b}\right], (a > 1, b > 1)$:

If the dimension is very high, we can assume that, compared with *r*, there are large numbers of eigenvalues on both side of 1. Based on previous observations, we should choose the eigenvalues on only one side of 1. If Σ's eigenvalues are uniformly distributed, then compared with the interval of $[1, \mathrm{b}]$, in the interval of $[\frac{1}{a}, 1]$ there are fewer eigenvalues, and the eigenvalues' "utility" for TV decrease faster (For example, eigenvalues that are uniformly distributed in the interval of [0.1, 0.4], are not likely to be as good as eigenvalues that are uniformly distributed in the interval of [9.7, 10.0]). So, if our observations in two dimensional cases can be generalized to higher dimension, then I suggest that we choose all the eigenvalues that are larger than 1 to maximize TV in these asymptotic cases.

5. Numerical Tests for several approximations for Total Variation:

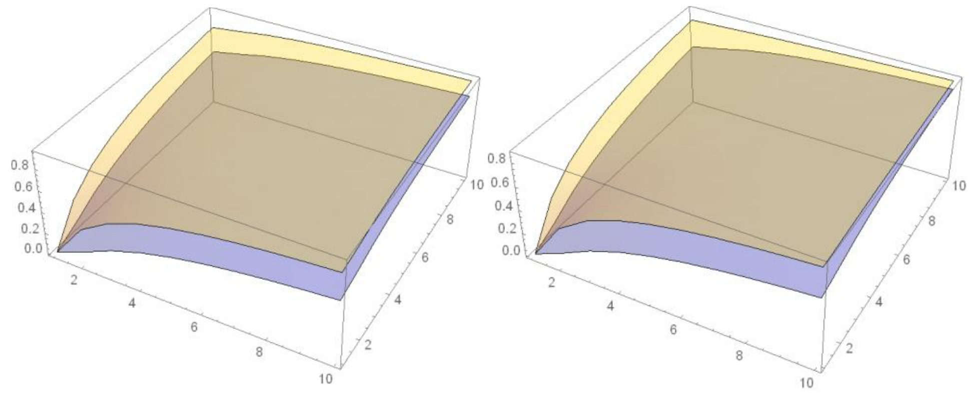
1) Total Variation Versus $H(P||Q)$:

Although $H(P||Q)$ seems to be a good approximation and a tight lower bound for total variation when $\sigma_1^2 > 1, \sigma_2^2 > 1$, it is no longer a lower bound for TV when $\sigma_1^2 > 1, \sigma_2^2 < 1$:



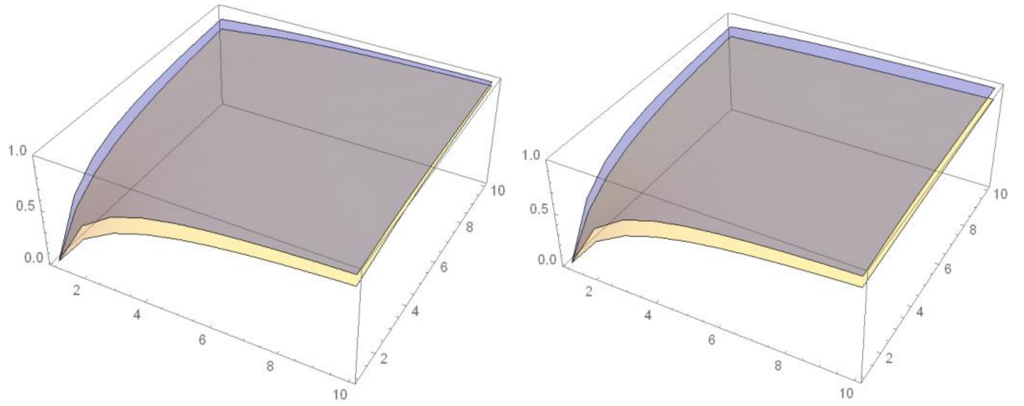
The left image is when $\sigma_1^2 > 1, \sigma_2^2 > 1$ and the right image is when $\sigma_1^2 > 1, \sigma_2^2 < 1$. The orange one is Total Variation and the blue one is Hellinger distance. These images once again show difference between Hellinger distance and TV: although they are numerically close, they have different properties. For Hellinger distance, σ_2 and $\frac{1}{\sigma_2}$ are symmetric, but for total variation, they are not symmetric.

2) Total Variation Versus $H^2(P||Q)$:



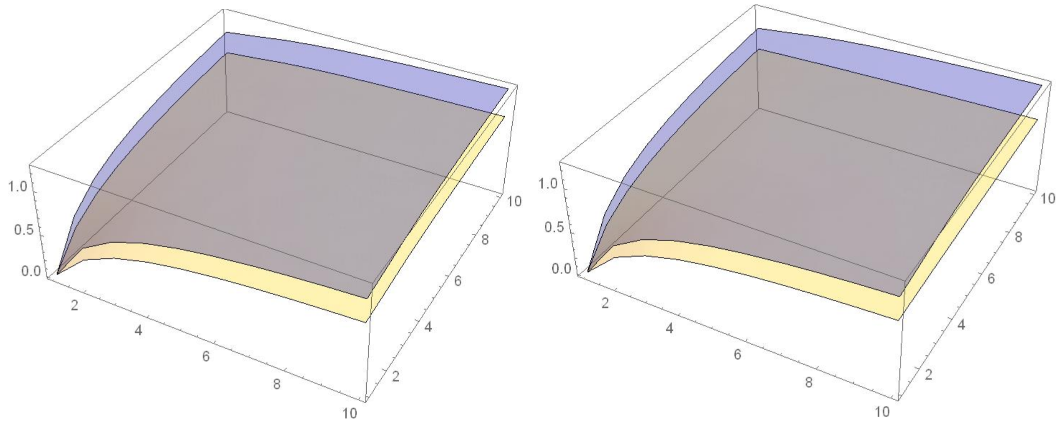
The left image is when $\sigma_1^2 > 1, \sigma_2^2 > 1$ and the right image is when $\sigma_1^2 > 1, \sigma_2^2 < 1$. The orange one is Total Variation and the blue one is $H^2(P||Q)$. These images confirm that TVs are lower bounded by $H^2(P||Q)$.

3) Total Variation Versus $\boldsymbol{H(P||Q)\sqrt{2 - H^2(P||Q)}}$:



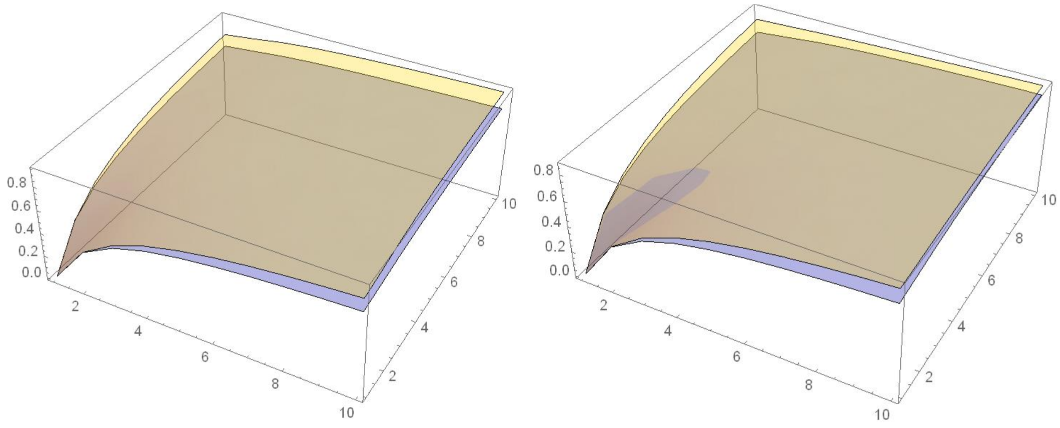
The left image is when $\sigma_1^2 > 1, \sigma_2^2 > 1$ and the right image is when $\sigma_1^2 > 1, \sigma_2^2 < 1$. The orange one is Total Variation and the blue one is $H(P||Q)\sqrt{2 - H^2(P||Q)}$. These images confirm that TVs are upper bounded by $H(P||Q)\sqrt{2 - H^2(P||Q)}$.

4) Total Variation Versus $\boldsymbol{\sqrt{2}H(P||Q)}$:



The left image is when $\sigma_1^2 > 1, \sigma_2^2 > 1$ and the right image is when $\sigma_1^2 > 1, \sigma_2^2 < 1$. The orange one is Total Variation and the blue one is $\sqrt{2}H(P||Q)$. These images confirm that TVs are upper bounded by $\sqrt{2}H(P||Q)$.

5) Total Variation Versus $\boldsymbol{H(P||Q)\sqrt{1 - \frac{H^2(P||Q)}{4}}}$:



The left image is when $\sigma_1^2 > 1, \sigma_2^2 > 1$ and the right image is when $\sigma_1^2 > 1, \sigma_2^2 < 1$. The orange one is Total Variation and the blue one is $H(P||Q)\sqrt{1-\frac{H^2(P||Q)}{4}}$. It seems that when $\sigma_1^2 > 1, \sigma_2^2 > 1$, TVs might be lower bounded by $H(P||Q)\sqrt{1-\frac{H^2(P||Q)}{4}}$, and when $\sigma_1^2 > 1, \sigma_2^2 < 1$, $H(P||Q)\sqrt{1-\frac{H^2(P||Q)}{4}}$ is neither an upper bound nor a lower bound for TVs.