The Comparison Among *f*-divergences, and some Robustness Analysis of KL-divergence and Hellinger Distance in High Dimension

1. Comparison among *f*-divergences:

The statistical distance, KL-divergence, Hellinger distance, and total variation, are all *f*-divergences:

$$D_f(p||q) = \int q(x) f(\frac{p(x)}{q(x)}) dx$$

in which *f* is a convex function and $f(1) = 0$.

As the total variation has no closed-form for Gaussian distribution, we alternatively use the following:

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

to estimate total variation (In the following discussion, I will call this the equivalent of total variation). In the above expression, $\beta_1, \beta_2, \dots, \beta_r$ are eigenvalues of $(AA^T)^{-1}Z - I_r$.

From previous analysis we have learned that, to maximize KL-divergence, Hellinger distance, and the equivalent of total variation, is to maximize some matrix functions of AA^T and $Z = A\Sigma A^T$.

Specifically, to maximize KL-divergence is to maximize the following:

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

To maximize Hellinger distance is to maximize the following:

$$\sqrt{1 - \frac{|A\Sigma A^T|^{\frac{1}{4}} |AA^T|^{\frac{1}{4}}}{\left|\frac{A\Sigma A^T + AA^T}{2}\right|^{\frac{1}{2}}}}$$

Which is equivalent to maximizing the following:

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

To maximize the equivalent of total variation is to maximize the following:

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

It is easy to verify that **all the three objective functions are invariant under linear transformations on the row vectors of *A***. So we can always assume that $AA^T = I_r$.

Hence, our problems reduce to find semi-orthogonal matrices $A \in \mathbb{R}^{r \times n}$ so that $Z = A\Sigma A^T$ is a compression of Σ and $\log|Z| + tr[Z^{-1}]$, $\frac{|Z+I_r|^2}{|Z|}$, $tr[(Z - I_r)^2]$ reach the maxima.

So how to solve these problems? The general idea is to project Σ onto its linear subspaces spanned by certain eigenvectors of Σ. By similar analysis applying Cauchy interlacing theorem (or Poincare separation theorem), we have discovered that we should identify *r* eigenvalues of Σ that deviate the most from 1 in some sense.

Specifically, to maximize KL-divergence, we should find *r* eigenvalues of Σ, $\lambda_1, \dots, \lambda_r$, so that:

$$\log(\lambda_i) + \frac{1}{\lambda_i} (1 \le i \le r)$$

are the *r* largest among all the eigenvalues.

To maximize Hellinger distance, we should find *r* eigenvalues of Σ, $\lambda_1, \dots, \lambda_r$, so that:

$$\lambda_i + \frac{1}{\lambda_i} (1 \le i \le r)$$

are the *r* largest among all the eigenvalues.

Similarly, to maximize the equivalent of total variation, we should find *r* eigenvalues of Σ, $\lambda_1, \dots, \lambda_r$, so that:

$$(\lambda_i - 1)^2 (1 \le i \le r)$$

are the *r* largest among all the eigenvalues.

Then we are able to construct *A* by the corresponding eigenvectors. By doing so, we compress *A* and project *A* onto a linear subspace, preserving all the most important *r* components.

There is only a slight difference among the three *f*-divergences. That is, by choosing different divergences we select the eigenvalues of Σ a little bit differently. If we use Hellinger distance, for example, we equivalently favor $\lambda(\lambda > 1)$ and $\frac{1}{\lambda}$. If we choose KL-divergence, it means that we favor the smaller ones, $\frac{1}{\lambda}(\lambda > 1)$, over the larger ones, $\lambda(\lambda > 1)$. If we use the equivalent of total variation, on the contrary, we favor the larger ones, $\lambda(\lambda > 1)$, over the smaller ones, $\frac{1}{\lambda}(\lambda > 1)$.

2. Robustness analysis of KL-divergence and Hellinger distance in high dimension:

Our aim is to detect the change of the signals with minimal delay and minimal false alarms. We argue that, in reality, even if two observations of signals are derived from the same statistical model, the covariance matrices are likely to vary because of approximations in calculations and noise in environment. Because of the perturbation, the covariance matrices are not likely to be exactly the same. However, we can show that the small errors can add up to a massive difference if we measure the statistical distance by KL-divergence or Hellinger distance in very high dimension.

First, we analysis Hellinger distance.

We have learned that Hellinger distance is:

$$\sqrt{1 - \prod_{i=1}^{r} (2^{\frac{1}{2}} \frac{\lambda_i^{\frac{1}{4}}}{(\lambda_i + 1)^{\frac{1}{2}}})}$$

When $\lambda_i = 1$, the factor $2^{\frac{1}{2}} \frac{\lambda_i^{\frac{1}{4}}}{(\lambda_i+1)^{\frac{1}{2}}}$ equals 1. If any eigenvalue λ_i changes from 1 to $1 - \Delta x$ because of perturbation, then $2^{\frac{1}{2}} \frac{\lambda_i^{\frac{1}{4}}}{(\lambda_i+1)^{\frac{1}{2}}}$ changes to $\left(1 - \frac{\Delta x}{2-\Delta x}\right)^{\frac{1}{4}} \left(1 + \frac{\Delta x}{2-\Delta x}\right)^{\frac{1}{4}}$.

Use the Taylor's series:

$$(1+x)^{\frac{1}{4}} = 1 + \frac{x}{4} - \frac{3}{32}x^2 + o(x^2)$$

We have:

$$\left(1 - \frac{\Delta x}{2-\Delta x}\right)^{\frac{1}{4}} \left(1 + \frac{\Delta x}{2-\Delta x}\right)^{\frac{1}{4}} = 1 - \frac{1}{4}\frac{(\Delta x)^2}{(2-\Delta x)^2} + o(\Delta x^2) \approx 1 - \frac{1}{16}(\Delta x)^2$$

If dimension is *N* and each λ_i changes from 1 to $1-\Delta x$, then:

$$\prod_{i=1}^{N} (2^{\frac{1}{2}} \frac{\lambda_i^{\frac{1}{4}}}{(\lambda_i+1)^{\frac{1}{2}}}) \approx e^{\sum_{i=1}^{N} \log\left[1-\frac{1}{16}(\Delta x)^2\right]} \approx e^{-\frac{N}{16}(\Delta x)^2}$$

So, if dimension *N* is very high, then $e^{-\frac{N}{16}(\Delta x)^2}$ will be small and Hellinger distance will be close to 1, which means that the two distributions will be considered very different by Hellinger distance even if each eigenvalue is similar.

For example, if $\Delta x = 0.01$, *N*=160,000, then $\prod_{i=1}^{N}(2^{\frac{1}{2}} \frac{\lambda_i^{\frac{1}{4}}}{(\lambda_i+1)^{\frac{1}{2}}}) \approx 0.364187$, which is very close to $\frac{1}{\mathrm{e}} \approx 0.367879$. In this case Hellinger distance is 0.797379, indicating that the two distributions, I_N and $0.99I_N$, are very different.

The above discussion means that, the small error in each dimension can add up to a very large difference in high dimension, and Hellinger distance might be not a very robust estimator of statistical distance in high dimensional cases.

One might expect that KL-divergence is more robust, since the value of KL-divergence is ranged from 0 to infinity, whereas the value of Hellinger distance is confined to the interval of $[0,1]$. However, we discovered that this is only partly true. In certain high dimensional cases, KL-divergence might not be a robust and practical estimator of statistical distance either.

If the dimension is *N*, then KL-divergence is:

$$\frac{1}{2}\sum_{i=1}^{N}(\log\lambda_i + \frac{1}{\lambda_i} - 1)$$

By Taylor's series:

$$\log(1+\Delta x) = \Delta x - \frac{(\Delta x)^2}{2} + o(\Delta x^2)$$

$$\frac{1}{1+\Delta x} = 1 - \Delta x + (\Delta x)^2 + \mathrm{o}(\Delta x^2)$$

We can deduce that each eigenvalue's increasing from 1 to $1+\Delta x$ means an increment of KL-divergence by $\frac{(\Delta x)^2}{4}$. As dimension is *N*, we deduce that the KL-divergence will increase by $\frac{\mathrm{N}}{4}(\Delta x)^2$ if every eigenvalue increases from 1 to $1+\Delta x$.

Since KL-divergence is ranged from 0 to infinity, it seems that $\frac{\mathrm{N}}{4}(\Delta x)^2$ is not very large. This is true, if some of the eigenvalues of Σ is very close to 0, since $\log\lambda + \frac{1}{\lambda} - 1 \rightarrow \infty$ as $\lambda \rightarrow 0$. However, we also notice that $\log\lambda + \frac{1}{\lambda} - 1$ grows rather slow in the interval of $[1, +\infty)$, so $\frac{N}{4}(\Delta x)^2$ is still comparable to Σ's some large eigenvalue's contribution to KL-divergence. In that case, **we can't distinguish one anormaly from the sum of a lot of small noise** by KL-divergence. So I think KL-divergence might also be not very robust in high dimensional cases in some sense.

3. The advantage of dimension reduction:

One of the advantages of dimension reduction besides computational efficiency, is that it helps us distinguish whether the difference measured by statistical distance is attributed to the sum of a lot of small noise, or is attributed to some major changes.

Under the "sparsity" assumption, dimension reduction might be even more appealing. If we assume that the changes are low rank in their nature, then this dimension reduction method will effectively identify the changes without making much computational efforts.

4. Some preliminary numerical results in the asymptotic aspect:

Assume that dimension $N \rightarrow \infty$ and the eigenvalues of Σ are evenly distributed in the interval of $[a, b]$, then we can alternatively use the following to estimate the KL-divergence:

$$\frac{1}{2}\int_{a}^{b}\left(log\lambda + \frac{1}{\lambda} - 1\right)d\lambda$$

If $a = 0.1$ and $b = 10$, then the "top" 10% eigenvalues contribute to approximately 23.5% of KL-divergence; If $a = 0.01$ and $b = 100$, then the "top" 10% eigenvalues contribute to approximately 14.0% of KL-divergence.