Gradient-based Optimization for Maximizing Total Variation

Sihui Wang

1. A Generalization of Total Variation: *k*-Generalized Total Variation

1.1 Definition

Suppose there are 2 P.D.F.s, $f(x)$ and $g(x)$, then we define the *k-generalized total variation* as the following:

$$D_{k-TV}(f(x)\|g(x)) = \int_{\mathbb{R}^n} |f(x) - kg(x)|dx$$

If $k = 1$, then the k-generalized total variation will be reduced to the total variation in the ordinary sense.

1.2 Motivation

Consider a random variable X. Suppose that $X \sim X_1$ with the probability of p, and $X \sim X_2$ with the probability of $1 - p$. Suppose that X_1's pdf is $f(x)$ and X_2's pdf is $g(x)$.

Now we sample from the random variable X and obtain x, and we are making the decision whether x is from X_1 or X_2 based on the likelihood ratio principle. If $f(x) > mg(x)$, then we decide that x is from X_1; if $f(x) \leq mg(x)$, then we decide that x is from X_2.

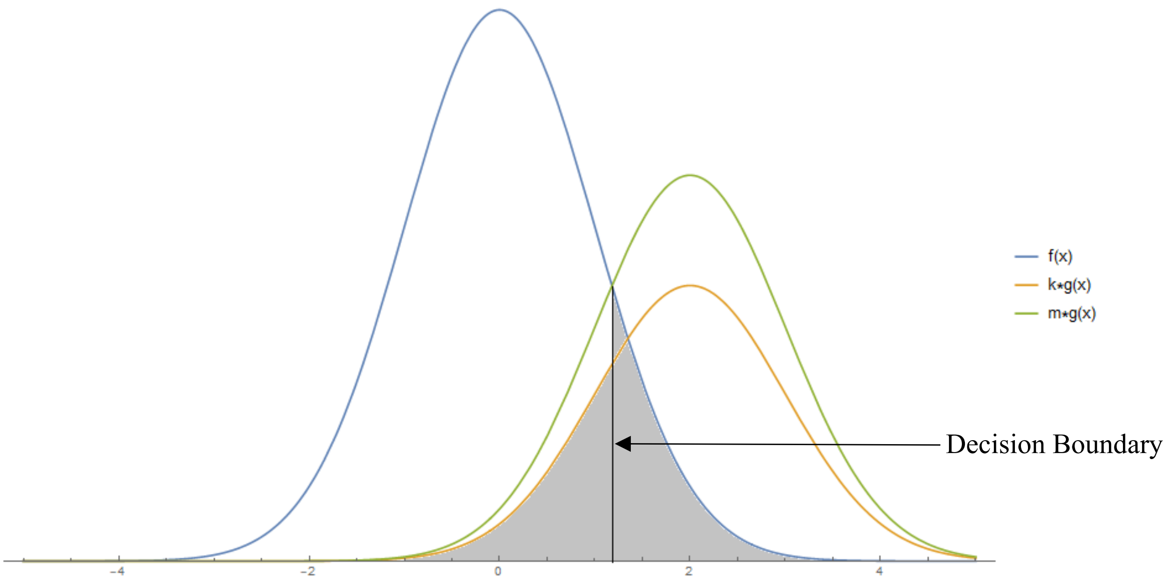
Given such a decision rule, the overall misclassification rate $\boldsymbol{P_{mis}}$ is:

$$\boldsymbol{P_{mis}} = p \cdot \mathbf{P}\{f(x) \leq mg(x) | x \in X_1\} + (1-p) \cdot \mathbf{P}\{f(x) > mg(x) | x \in X_2\}$$

Denoting $\frac{1-p}{p}$ by k, it is easy to confirm that minimizing $\boldsymbol{P_{mis}}$ is equivalent to minimizing the following:

$$\frac{1}{p}\boldsymbol{P_{mis}} = \mathbf{P}\{f(x) \leq mg(x) | x \in X_1\} + k \cdot \mathbf{P}\{f(x) > mg(x) | x \in X_2\}$$

It is easy to confirm that $\frac{1}{p}\boldsymbol{P_{mis}}$ is the area of the shadow region in the figure below:



Roughly speaking, $\frac{1}{p}\boldsymbol{P_{mis}}$ is the sum of $f(x)$'s "tail" in the right and $kg(x)$'s "tail" in the left. When the dataset is given, p and $k=\frac{1-p}{p}$ are fixed, and the only variable is m that we have for the decision rule. Note in the figure above that the decision boundary will move as m, the parameter for our decision rule, changes. It is easy to confirm that $\frac{1}{p}\boldsymbol{P_{min}}$ will be minimized when $m=k$.

When $m=k$, we have the following equation for the minimized $\frac{1}{p}\boldsymbol{P_{mis}}$:

$$\left(\frac{1}{p}\boldsymbol{P_{mis}}\right)_{min}=\frac{1}{2}\left(k+1-\int_{\mathbb{R}^{\mathrm{n}}}|f(x)-kg(x)|dx\right)$$

or:

$$(\boldsymbol{P_{mis}})_{min}=\frac{1}{2}\left(1-p\int_{\mathbb{R}^{\mathrm{n}}}|f(x)-kg(x)|dx\right)=\frac{1}{2}\left(1-\int_{\mathbb{R}^{\mathrm{n}}}|pf(x)-(1-p)g(x)|dx\right)$$

Note that we have defined that

$$D_{k-TV}(f(x)\|g(x))=\int_{\mathbb{R}^{\mathrm{n}}}|f(x)-kg(x)|dx$$

So **minimizing the misclassification rate $\boldsymbol{P_{mis}}$ is equivalent to maximizing the *k*-*generalized total variation*, $\boldsymbol{D_{k-TV}(f(x)\|g(x))}$**: this is the motivation why we defined the *k*-*generalized total variation* in the first place.

2. Differentiating the *k*-Generalized Total Variation

2.1 Variational Analysis in General Cases

Now let us consider the following problem: if we add a small perturbation to $g(x)$ and obtain $\tilde{g}(x)=g(x)+\varepsilon\eta(x)$, how should we quantify $D_{k-TV}(f(x)\|\tilde{g}(x))-D_{k-TV}(f(x)\|g(x))$?

In fact, we have the following:

$$\begin{aligned}D_{k-TV}(f(x)\|\tilde{g}(x))-D_{k-TV}(f(x)\|g(x))&=\int_{\mathbb{R}^{\mathrm{n}}}|f(x)-k\tilde{g}(x)|dx-\int_{\mathbb{R}^{\mathrm{n}}}|f(x)-kg(x)|dx\\&=\int_{f(x)\geq k\tilde{g}(x)}[f(x)-k\tilde{g}(x)]dx+\int_{f(x)<k\tilde{g}(x)}[k\tilde{g}(x)-f(x)]dx\\&-\int_{f(x)\geq kg(x)}[f(x)-kg(x)]dx-\int_{f(x)<kg(x)}[kg(x)-f(x)]dx\\&=k[\int_{S_1}(\tilde{g}(x)-g(x))dx+\int_{S_2}(g(x)-\tilde{g}(x))dx)]\\&+\int_{S_3}[k(\tilde{g}(x)+g(x))-2f(x)]dx+\int_{S_4}[2f(x)-k(\tilde{g}(x)+g(x))]dx\end{aligned}$$

where

$S_1=\{x\in\mathbb{R}^{\mathrm{n}}|f(x)<min\{k\tilde{g}(x),kg(x)\}\}$, $S_2=\{x\in\mathbb{R}^{\mathrm{n}}|f(x)>max\{k\tilde{g}(x),kg(x)\}\}$,
$S_3=\{x\in\mathbb{R}^{\mathrm{n}}|k\tilde{g}(x)>f(x)>kg(x)\}$, $S_4=\{x\in\mathbb{R}^{\mathrm{n}}|kg(x)>f(x)>k\tilde{g}(x)\}$.

Given $\varepsilon\to 0$, we have the following:

$$\int_{S_3}[k(\tilde{g}(x)+g(x))-2f(x)]dx=O(\varepsilon^2)$$

Intuitively, this is because the Lebesgue measure of S_3 is $O(\varepsilon)$, and $k(\tilde{g}(x)+g(x))-2f(x)$ is $O(\varepsilon)$ on $S_3=\{x\in\mathbb{R}^{\mathrm{n}}|k\tilde{g}(x)>f(x)>kg(x)\}$ as $\varepsilon\to 0$.

For the same reason, we have the following as $\varepsilon\to 0$:

$$\int_{S_4} [2f(x) - k(\tilde{g}(x) + g(x))]dx = O(\varepsilon^2)$$

In addition, we have:

$$\int_{S_1} (\tilde{g}(x) - g(x))dx = \int_{f(x)<kg(x)} (\tilde{g}(x) - g(x))dx + O(\varepsilon^2)$$

$$\int_{S_2} (g(x) - \tilde{g}(x))dx) = \int_{f(x)>kg(x)} (g(x) - \tilde{g}(x))dx + O(\varepsilon^2)$$

To see this, we just need to verify that:

$$\int_{S_1} (\tilde{g}(x) - g(x))dx - \int_{f(x)<kg(x)} (\tilde{g}(x) - g(x))dx$$

$$= -\int_{k\tilde{g}(x)<f(x)<kg(x)} (\tilde{g}(x) - g(x))dx = O(\varepsilon^2)$$

and

$$\int_{S_2} (g(x) - \tilde{g}(x))dx) - \int_{f(x)>kg(x)} (g(x) - \tilde{g}(x))dx$$

$$= -\int_{kg(x)<f(x)<k\tilde{g}(x)} (g(x) - \tilde{g}(x))dx = O(\varepsilon^2)$$

This is because the Lebesgue measure of $\{x \in \mathbb{R}^{\mathbf{n}} | k\tilde{g}(x) < f(x) < kg(x)\}$ and $\{x \in \mathbb{R}^{\mathbf{n}} | kg(x) < f(x) < k\tilde{g}(x)\}$ are $O(\varepsilon)$ and $\tilde{g}(x) - g(x) = O(\varepsilon)$.

Therefore, we have:

$$D_{k-TV}(f(x)\|\tilde{g}(x)) - D_{k-TV}(f(x)\|g(x))$$
$$= k[\int_{f(x)<kg(x)} (\tilde{g}(x) - g(x))dx + \int_{f(x)>kg(x)} (g(x) - \tilde{g}(x))dx] + O(\varepsilon^2)$$

Noting that

$$\int_{f(x)<kg(x)} (\tilde{g}(x) - g(x))dx + \int_{f(x)>kg(x)} (\tilde{g}(x) - g(x))dx = \int_{\mathbb{R}^{\mathbf{n}}} \tilde{g}(x)dx - \int_{\mathbb{R}^{\mathbf{n}}} g(x)dx = 0$$

we finally have the following:

$$D_{k-TV}(f(x)\|\tilde{g}(x)) - D_{k-TV}(f(x)\|g(x)) = 2k \int_{f(x)<kg(x)} (\tilde{g}(x) - g(x))dx + O(\varepsilon^2)$$

Or:

$$D_{k-TV}(f(x)\|\tilde{g}(x)) - D_{k-TV}(f(x)\|g(x)) = -2k \int_{f(x)>kg(x)} (\tilde{g}(x) - g(x))dx + O(\varepsilon^2)$$

2.2 Specific Analysis for the Cases of Normal Distributions

Now suppose $g(x)$ is the P.D.F of the normal distribution $\mathcal{N}(\mu, \Sigma)$ and $\tilde{g}(x)$ is the P.D.F. of the normal distribution $\mathcal{N}(\mu', \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_i, \ldots, \mu_n)$, $\mu' = (\mu_1, \ldots, \mu_i + \Delta\mu_i, \ldots, \mu_n)$, and $\Sigma = (\sigma_{ij})(1 \le i, j \le n)$.

Then we have:

$$D_{k-TV}(f(x)\|\tilde{g}(x)) - D_{k-TV}(f(x)\|g(x)) = 2k \int_{f(x)<kg(x)} (\tilde{g}(x) - g(x))dx + O((\Delta\mu_i)^2)$$

Hence, we have:

$$\frac{\partial D_{k-TV}(f\|g)}{\partial \mu_i} = 2k \int_{f<kg} \frac{\partial g}{\partial \mu_i} dx$$

Similarly, we have:

$$\frac{\partial D_{k-TV}(f\|g)}{\partial \sigma_{ij}} = 2k\int_{f<kg}\frac{\partial g}{\partial \sigma_{ij}}dx$$

Or equivalently:

$$\frac{\partial D_{k-TV}(f\|g)}{\partial \mu_i} = -2k\int_{f>kg}\frac{\partial g}{\partial \mu_i}dx$$

$$\frac{\partial D_{k-TV}(f\|g)}{\partial \sigma_{ij}} = -2k\int_{f>kg}\frac{\partial g}{\partial \sigma_{ij}}dx$$

2.3 Linear Dimensionality Reduction for a Mixture of two Gaussians:

Now suppose $X \in \mathbb{R}^n$ is a mixture of two Gaussians. With probability p, $X = X_1 \sim \mathcal{N}(0, I_n)$ and with probability $1-p$, $X = X_2 \sim \mathcal{N}(\mu, \Sigma)$. Suppose we impose on X a linear transformation $A \in \mathbb{R}^{r\times n}(AA^T = I_r)$ and obtain AX. Then with probability p, $AX = AX_1 \sim \mathcal{N}(0, I_r)$ and with probability $1-p$, $AX = AX_2 \sim \mathcal{N}(A\mu, A\Sigma A^T)$. Our goal is to find an optimal A to minimize the misclassification rate P_{mis}.

From the above discussion, we know that minimizing P_{mis} is equivalent to maximizing the following:

$$D_{k-TV}(AX_1\|AX_2)$$

So, our objective is to:

$$\text{maximize } D_{k-TV}(AX_1\|AX_2)$$
$$s.t.\ A \in \mathbb{R}^{r\times n}, AA^T = I_r$$

Note that under the constraint $AA^T = I_r$, it is guaranteed that $AX_1 \sim \mathcal{N}(0, I_r)$, so our objective could be rewritten as the following:

$$\text{maximize } D_{k-TV}(\mathcal{N}(0, I_r)\|\mathcal{N}(A\mu, A\Sigma A^T))$$
$$s.t.\ A \in \mathbb{R}^{r\times n}, AA^T = I_r$$

Suppose that $f(x)$ is the P.D.F. of $\mathcal{N}(0, I_r)$, and $g(x)$ is the P.D.F. of $\mathcal{N}(A\mu, A\Sigma A^T)$ where $A\mu = (\mu_1, \dots, \mu_r)$ and $A\Sigma A^T = (\sigma_{ij})$. Note that we have already had the formulas:

$$\frac{\partial D_{k-TV}(f\|g)}{\partial \mu_i} = 2k\int_{f<kg}\frac{\partial g}{\partial \mu_i}dx = -2k\int_{f>kg}\frac{\partial g}{\partial \mu_i}dx$$

$$\frac{\partial D_{k-TV}(f\|g)}{\partial \sigma_{ij}} = 2k\int_{f<kg}\frac{\partial g}{\partial \sigma_{ij}}dx = -2k\int_{f>kg}\frac{\partial g}{\partial \sigma_{ij}}dx$$

So, by chain rule we have:

$$\nabla_A D_{k-TV}(f\|g) = \sum_i \frac{\partial D_{k-TV}(f\|g)}{\partial \mu_i}\cdot\frac{\partial \mu_i}{\partial A} + \sum_{i,j}\frac{\partial D_{k-TV}(f\|g)}{\partial \sigma_{ij}}\cdot\frac{\partial \sigma_{ij}}{\partial A}$$

Hence, we obtained an expression for $\nabla_A D_{k-TV}(f\|g)$. So, to find the optimal solution for linear dimensionality reduction with the minimized misclassification rate P_{mis} is equivalent to find A so that $\nabla_A D_{k-TV}(f\|g) = 0$, which could be solved by gradient descent algorithms.

2.4 The Special Case of $r = 1$:

Denoting $A\mu$ by μ' and $A\Sigma A^T$ by σ'^2, we have:

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

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{(x-\mu')^2}{2\sigma'^2}}$$

where $f(x)$ is the P.D.F. of $\mathcal{N}(0,1)$, and $g(x)$ is the P.D.F. of $\mathcal{N}(A\mu, A\Sigma A^T)$. Our objective is to find an optimal $A \in \mathbb{R}^{1\times n}$ so that the k-generalized total variation between $f(x)$ and $g(x)$

is maximized.

$$\nabla_A D_{k-TV}(f\|g) = 2k\int_{f<kg}\frac{\partial g}{\partial \mu'}dx\cdot\frac{\partial \mu'}{\partial A} + 2k\int_{f<kg}\frac{\partial g}{\partial \sigma'}dx\cdot\frac{\partial \sigma'}{\partial A}$$

$$= 2k\left[\int_{f<kg}\frac{\partial g}{\partial \mu'}dx\cdot\mu^T + \int_{f<kg}\frac{\partial g}{\partial \sigma'}dx\cdot\frac{\partial\sqrt{A\Sigma A^T}}{\partial A}\right]$$

$$= 2k\left[\int_{f<kg}\frac{\partial g}{\partial \mu'}dx\cdot\mu^T + \int_{f<kg}\frac{\partial g}{\partial \sigma'}dx\cdot\frac{A\Sigma}{\sigma'}\right]$$

$$= 2k\left[\int_{f<kg}\frac{x-\mu'}{\sqrt{2\pi}\sigma'^3}e^{-\frac{(x-\mu')^2}{2\sigma^2}}dx\cdot\mu^T\right.$$

$$\left.+\int_{f<kg}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu')^2}{2\sigma'^2}}\left(\frac{(x-\mu')^2}{\sigma'^4}-\frac{1}{\sigma'^2}\right)dx\cdot\frac{A\Sigma}{\sigma'}\right]$$

Let $\tilde{x} = \frac{x-\mu\prime}{\sigma\prime}$, then we have:

$$\nabla_A D_{k-TV}(f\|g) = 2k\left[\int_{f<kg}\frac{\tilde{x}}{\sqrt{2\pi}\sigma'}e^{-\frac{\tilde{x}^2}{2}}d\tilde{x}\ \cdot\mu^T + \int_{f<kg}\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{\tilde{x}^2}{2}}(\tilde{x}^2-1)d\tilde{x}\cdot\frac{A\Sigma}{\sigma'}\right]$$

Or:

$$\nabla_A D_{k-TV}(f\|g) = -2k\left[\int_{f>kg}\frac{\tilde{x}}{\sqrt{2\pi}\sigma'}e^{-\frac{\tilde{x}^2}{2}}d\tilde{x}\ \cdot\mu^T + \int_{f>kg}\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{\tilde{x}^2}{2}}(\tilde{x}^2-1)d\tilde{x}\cdot\frac{A\Sigma}{\sigma'}\right]$$

So, basically the ***k*-generalized total variation is determined by the integrals of $\boldsymbol{\phi_1(x) = \frac{1}{\sqrt{2\pi}}xe^{-\frac{x^2}{2}}}$ and $\boldsymbol{\phi_2(x) = \frac{1}{\sqrt{2\pi}}(x^2-1)e^{-\frac{x^2}{2}}}$ on the region of $\boldsymbol{f<kg}$ or $\boldsymbol{f>kg}$**.

When $r = 1$, it is easy to find out the interval(s) this is (are) corresponding to $f < kg$, and we can maintain a look-up table to find out the integrals of $\phi_1(x)$ and $\phi_2(x)$ on certain intervals. Hence, we are able to write down the closed-form of *k*-generalized total variation which is numerically obtainable when $r = 1$.

Deciding the Integral Region:

Since $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $g(x) = \frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{(x-\mu')^2}{2\sigma'^2}}$, we deduce that $f < kg$ is equivalent to the following:

$$-\frac{x^2}{2} < \log\left(\frac{k}{\sigma'}\right) - \frac{(x-\mu')^2}{2\sigma'^2}$$

which is equivalent to the following:

$$\frac{1}{2}\left(1-\frac{1}{\sigma'^2}\right)x^2 + \frac{\mu'}{\sigma'^2}x + \log\left(\frac{k}{\sigma'}\right) - \frac{\mu'^2}{2\sigma^2} > 0$$

If $\sigma' = 1$, then the above inequality becomes degenerate.

If $\sigma' = 1$, $\mu' = 0$, then we obtain $f(x) = g(x)$. So, we simply have:

$$D_{k-TV}(f\|g) = |k-1|$$
$$\nabla_A D_{k-TV}(f\|g) = 0$$

When $\sigma' = 1$, $\mu' > 0$, $f < kg$ is equivalent to:

$$x > \frac{\mu'}{2} - \frac{\log(k)}{\mu'}$$

$$\tilde{x} = \frac{x-\mu'}{\sigma'} > -\frac{\mu'}{2} - \frac{\log(k)}{\mu'}$$

$$\nabla_A D_{k-TV}(f||g) = \frac{2k}{\sigma'}\left[\int_{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}}^{+\infty} \phi_1(\tilde{x})d\tilde{x} \cdot \mu^T + \int_{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}}^{+\infty} \phi_2(\tilde{x})d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}\right]$$

When $\sigma' = 1,\ \mu' < 0,\ f < kg$ is equivalent to:

$$x < \frac{\mu'}{2} - \frac{\log(k)}{\mu'}$$

$$\tilde{x} = \frac{x-\mu'}{\sigma'} < -\frac{\mu'}{2} - \frac{\log(k)}{\mu'}$$

$$\nabla_A D_{k-TV}(f||g) = \frac{2k}{\sigma'}\left[\int_{-\infty}^{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}} \phi_1(\tilde{x})d\tilde{x} \cdot \mu^T + \int_{-\infty}^{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}} \phi_2(\tilde{x})d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}\right]$$

$$= -\frac{2k}{\sigma'}\left[\int_{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}}^{+\infty} \phi_1(\tilde{x})d\tilde{x} \cdot \mu^T + \int_{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}}^{+\infty} \phi_2(\tilde{x})d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}\right]$$

When $\sigma' \neq 1$, the constraint $\{x|f(x) < kg(x)\} = \{x|\frac{1}{2}\left(1-\frac{1}{\sigma'^2}\right)x^2 + \frac{\mu'}{\sigma'^2}x + \log\left(\frac{k}{\sigma'}\right) - \frac{\mu'^2}{2\sigma^2} > 0\}$ is quadratic and we need to check the discriminant first.

If $\sigma'^2\Delta = \mu'^2 + 2(1-\sigma'^2)\log\left(\frac{k}{\sigma'}\right) \leq 0$ and $\sigma' \neq 1$, we have:

$$D_{k-TV}(f||g) = |k-1|$$
$$\nabla_A D_{k-TV}(f||g) = 0$$

The interpretation is that either the curve of $f(x)$ is under the curve of $kg(x)$, or the curve of $kg(x)$ is under the curve of $f(x)$ in this case.

If $\sigma'^2\Delta = \mu'^2 + 2(1-\sigma'^2)\log\left(\frac{k}{\sigma'}\right) > 0$ and $\sigma' > 1$, we have $\{x|f(x) > kg(x)\} = \left\{x\middle|\frac{-\mu'-\sigma'^2\sqrt{\Delta}}{\sigma'^2-1} < x < \frac{-\mu'+\sigma'^2\sqrt{\Delta}}{\sigma'^2-1}\right\} = \{\tilde{x}|\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1} < \tilde{x} < \frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}\}$

$$\nabla_A D_{k-TV}(f||g) = -\frac{2k}{\sigma'}\left[\int_{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}}^{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}} \phi_1(\tilde{x})d\tilde{x} \cdot \mu^T + \int_{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}}^{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}} \phi_2(\tilde{x})d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}\right]$$

$$= \frac{2k}{\sigma'}\left[\int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_1(\tilde{x})d\tilde{x} \cdot \mu^T + \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_2(\tilde{x})d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}\right]$$

If $\sigma'^2\Delta = \mu'^2 + 2(1-\sigma'^2)\log\left(\frac{k}{\sigma'}\right) > 0$ and $\sigma' < 1$, we have $\{x|f(x) < kg(x)\} = \left\{x\middle|\frac{-\mu'+\sigma'^2\sqrt{\Delta}}{\sigma'^2-1} < x < \frac{-\mu'-\sigma'^2\sqrt{\Delta}}{\sigma'^2-1}\right\} = \{\tilde{x}|\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1} < \tilde{x} < \frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}\}$

$$\nabla_A D_{k-TV}(f||g) = \frac{2k}{\sigma'}\left[\int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_1(\tilde{x})d\tilde{x} \cdot \mu^T + \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_2(\tilde{x})d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}\right]$$

Closed-Form of the Gradient of *k*-Generalized Total Variation When $r = 1$:

To summarize, we have the following:

$$\nabla_A D_{k-TV}(f||g) = \begin{cases} \frac{sgn(\mu')2k}{\sigma'}\left[\int_{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}}^{+\infty}\phi_1(x)d\tilde{x}\cdot\mu^T + \int_{-\frac{\mu'}{2}-\frac{\log(k)}{\mu'}}^{+\infty}\phi_2(\tilde{x})d\tilde{x}\cdot\frac{A\Sigma}{\sigma'}\right] & \sigma'=1 \\ \max(0, sgn(\Delta))\frac{2k}{\sigma'}\left[\int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}}\phi_1(\tilde{x})d\tilde{x}\cdot\mu^T + \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}}\phi_2(\tilde{x})d\tilde{x}\cdot\frac{A\Sigma}{\sigma'}\right] & \sigma'\neq 1 \end{cases}$$

Let $k = 1$, we can also obtain the gradient of total variation:

$$\nabla_A D_{TV}(f||g) = \begin{cases} \frac{2\cdot sgn(\mu')}{\sigma'}\left[\int_{-\frac{\mu'}{2}}^{+\infty}\phi_1(x)d\tilde{x}\cdot\mu^T + \int_{-\frac{\mu'}{2}}^{+\infty}\phi_2(\tilde{x})d\tilde{x}\cdot\frac{A\Sigma}{\sigma'}\right] & \sigma'=1 \\ \max(0, sgn(\Delta))\frac{2}{\sigma'}\left[\int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}}\phi_1(\tilde{x})d\tilde{x}\cdot\mu^T + \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}}\phi_2(\tilde{x})d\tilde{x}\cdot\frac{A\Sigma}{\sigma'}\right] & \sigma'\neq 1 \end{cases}$$

where $\Delta = \frac{\mu'^2-2(1-\sigma'^2)\log(\sigma')}{\sigma'^2}$

Gradient-Based Optimization Algorithm ($r = 1$):

Given any $A \in \mathbb{R}^{1\times n}$, we can obtain $\mu' = A\mu$ and $\sigma' = \sqrt{A\Sigma A^T}$. Based on previous discussion, we can obtain $\nabla_A D_{k-TV}$ and update A accordingly, as long as we have a look-up table for the integrals of $\phi_1(x) = \frac{1}{\sqrt{2\pi}}xe^{-\frac{x^2}{2}}$ and $\phi_2(x) = \frac{1}{\sqrt{2\pi}}(x^2-1)e^{-\frac{x^2}{2}}$.

Gradient-Based Algorithm for Maximizing Total Variation ($r = 1$)

Input: $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$, k

Output: $A \in \mathbb{R}^{1\times n}$

1.Initialization: Randomly generate $A \in \mathbb{R}^{1\times n}$ so that $AA^T = 1$;

2.Iterations:

Do

<2.1> Calculate $\mu' = A\mu$, $\sigma' = \sqrt{A\Sigma A^T}$;

<2.2> Calculate $\nabla_A D_{k-TV}$: this is the gradient of k-generalized total variation in Euclidean space;

<2.3> Calculate $\nabla_A D_{k-TV} \leftarrow \nabla_A D_{k-TV} - <\nabla_A D_{k-TV}, A> A$: this is the gradient of k-generalized total variation on the manifold of $AA^T = 1$;

<2.4> Update $A \leftarrow A - \alpha\nabla_A D_{k-TV}$;

<2.5> Retraction $A \leftarrow \frac{A}{||A||}$ to make sure that $AA^T = 1$;

Until convergence;

3. Return A

Proportional Structure of Solutions:

Previously, we have observed that the one dimensional solutions for KL divergence, symmetric KL divergence, Hellinger's distance, and information geometric distance all show a "proportional structure". Here we prove that the one dimensional solutions for k-generalized total variation show exactly the same kind of proportional structures.

This comes from the fact that, at the critical point, $\nabla_A D_{k-TV}$ should be orthogonal to the manifold $AA^T = 1$, which means that:

$$\nabla_A D_{k-TV} = \gamma A$$

Based on previous discussions, we have:

$$\nabla_A D_{k-TV} = \frac{2k}{\sigma'} \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_1(\tilde{x}) d\tilde{x} \cdot \mu^T + \frac{2k}{\sigma'} \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_2(\tilde{x}) d\tilde{x} \cdot \frac{A\Sigma}{\sigma'}$$

Let $C_1 = \frac{2k}{\sigma'} \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_1(\tilde{x}) d\tilde{x}$,$C_2 = \frac{2k}{\sigma'^2} \int_{\frac{\sigma'(\sqrt{\Delta}-\mu')}{\sigma'^2-1}}^{\frac{-\sigma'(\sqrt{\Delta}+\mu')}{\sigma'^2-1}} \phi_2(\tilde{x}) d\tilde{x}$, we have:

$$\nabla_A D_{k-TV} = C_1 \cdot \mu^T + C_2 \cdot A\Sigma = \gamma A$$

Suppose the eigen-decomposition of Σ is $\Sigma = V\Lambda V^T$, μ and A have representations $\mu = V\alpha$, $A = \beta^T V^T$, then we have:

$$C_1 \alpha^T V^T + C_2 \beta^T V^T V \Sigma V^T = \gamma \beta^T V^T$$

Right-multiplying V on both sides, we obtain:

$$C_1 \alpha^T + C_2 \beta^T \Sigma = \gamma \beta^T$$

It is from this that we deduce that $\beta_1, \dots, \beta_n$ are proportional to each other:

$$\beta_1 : \dots : \beta_n = \frac{\alpha_1}{\lambda_1 + \gamma} : \dots : \frac{\alpha_n}{\lambda_n + \gamma}$$

Hence, we are able to explain that the one dimensional solutions for *k*-generalized total variation show exactly the same kind of proportional structures as KL divergence, symmetric KL divergence, Hellinger's distance, and information geometric distance do.

2.5 The General Cases when $r > 1$:

When $r > 1$, the integral region $f < kg$ becomes very complicated, which makes it increasingly difficult to obtain $\nabla_A D_{k-TV}$ numerically. However, we could resort to an approximation algorithm which iteratively find optimal one dimensional solutions and stack them together to formulate an approximated solution when $r > 1$. We have proposed similar algorithms for other *f*-divergence measures:

Greedy Algorithms for Linear Dimensionality Reduction

Input: $\mu_1, \mu_2, \Sigma_1, \Sigma_2, r$

Output: $A \in \mathbb{R}^{r \times n}$

1. Initialization. Solution vectors: $U \leftarrow \emptyset$;

Transformer: $T \leftarrow \Sigma_1^{-\frac{1}{2}}$; $\mu \leftarrow T(\mu_2 - \mu_1)$; $\Sigma \leftarrow T\Sigma_2 T$;

2.For *k*=1 to *r*

(1) Initialization step:

<1.1> Generate a non-zero vector *u*;

<1.2> Project *u* into the subspace: $u_{k,\perp} \leftarrow u - \sum_{i=1}^{k-1} < u, u_i > u_i$;

<1.3> Normalization: $u_{k,0} \leftarrow \frac{u_{k,\perp}}{\|u_{k,\perp}\|_2}$;

<1.4> $l \leftarrow 0$;

(2) **Do**

<2.1> Compute the gradient in Euclidean space: calculate $\nabla_{u_{k,l}} F$;

<2.2> Compute the gradient on the manifold: $\nabla^{M}_{u_{k,l}} F \leftarrow \nabla_{u_{k,l}} F - < \nabla_{u_{k,l}} F, u_{k,l} > u_{k,l}$;

<2.3> Gradient projection: $\nabla^{\perp}_{u_{k,l}} F \leftarrow \nabla^{M}_{u_{k,l}} F - \sum_{i=1}^{k-1} < \nabla^{M}_{u_{k,l}} F, u_i > u_i$;

<2.4> Update $u_{k,l}$ in the usual way: $u^{*}_{k,l} \leftarrow u_{k,l} + \alpha \nabla^{\perp}_{u_{k,l}} F$;

<2.5> Retraction: $u_{k,l+1} = \frac{u^{*}_{k,l}}{\|u^{*}_{k,l}\|_2}$

<2.6> $l \leftarrow l + 1$;

While not Convergence;

(3) Obtain $u_k \leftarrow u_{k,l}$ after the convergence of $u_{k,l}$; $U \leftarrow U \cup \{u_k\}$;

End For

3. $A^{*} \leftarrow (u_1^T, \dots, u_r^T)^T$;

Return $A \leftarrow A^{*}T$

With minor modification we can implement an approximation algorithm to find the maximized *k*-generalized total variation when $r > 1$.

3.Experiment:

Preliminary results can be found in Test_Result_TV.xlsx. For some datasets, the algorithm presented in this report produces results that are comparable to other algorithms that are aimed at maximizing other *f*-divergences. However, when the datasets are extremely unbalanced, the algorithm aimed at maximizing *k*-generalized total variation tends to produce worse outcomes.