An Information Geometric Interpretation of Linear Dimensionality Reduction

Sihui Wang

1.Preliminaries

Fisher's Information Matrices as Riemannian Metrics

Suppose there is a family of probability distributions, and their probabilistic density functions are $p(x;\xi)$ with the parameters $\xi = (\xi_1, \dots, \xi_n)$. Then *Fisher's Information Matrix*, $I(\xi)$, is defined as the following:

$$I(\xi)_{ij} = \mathbb{E}_{\xi}[\frac{\partial}{\partial \xi_i}\log(p(x;\xi))\frac{\partial}{\partial \xi_j}\log(p(x;\xi))]$$

In information geometry, we use Fisher's Information Matrices to define the Riemannian metrics on the statistical manifolds [1]:

$$g_{ij} = g\left(\frac{\partial}{\partial \xi_i}, \frac{\partial}{\partial \xi_j}\right) = I(\xi)_{ij}$$

With well-defined Riemannian metrics, we are able to embed a family of probability distributions in a manifold.

Connection and Geodesics

Since we have already defined Riemannian metrics on the manifolds, we can obtain *Levi-Civita connection* by *Christoffel symbols*:

$$\Gamma_{ij}^{k} = \frac{1}{2}\sum_{l} g^{kl}\left(\frac{\partial g_{il}}{\partial \xi_j} + \frac{\partial g_{lj}}{\partial \xi_i} - \frac{\partial g_{ij}}{\partial \xi_l}\right)$$

g^{kl} is the element in the *k*-th row and *l*-th column of the inverse matrix of $G = (g_{ij})$.

For Γ_{ij}^{k}, we have:

$$\nabla_{\frac{\partial}{\partial \xi_j}}\frac{\partial}{\partial \xi_i} = \sum_{k}\Gamma_{ij}^{k}\frac{\partial}{\partial \xi_k}$$

From Γ_{ij}^{k} we could obtain any covariant derivatives $\nabla_X Y$ ($X = \sum_i X_i \frac{\partial}{\partial \xi_i}$ and $Y = \sum_i Y_i \frac{\partial}{\partial \xi_i}$ are vector fields on the manifolds).

For geodesics $\gamma(t)$, we have $\nabla_{\gamma'(t)}\gamma'(t) = 0$, which could be rewritten as a system of differential equations:

$$\frac{d^2\xi_k(t)}{dt^2} + \sum_{i,j}\Gamma_{ji}^{k}(\gamma(t))\frac{d\xi_i(t)}{dt}\frac{d\xi_j(t)}{dt} = 0, 1 \le k \le n.$$

Hence, by defining Riemannian metric $G = (g_{ij})$, we are able to determine the geodesic $\gamma(t)$ whose length is:

$$\int_{t_1}^{t_2}\left[\left(\gamma'(t)\right)^T G\left(\gamma(t)\right)\gamma'(t)\right]^{\frac{1}{2}} dt$$

Fisher's information distance is defined as the length of the geodesic between two distributions, $p(x)$ and $q(x)$, on the statistical manifolds.

A manifold's geodesic could be different from the geodesic of its sub-manifold. For example, in $\mathbb{R}^3$ the geodesics are straight lines, while in the sub-manifold $\{(x,y,z)|x^2+y^2+z^2=1\}$ the geodesics are circles.

Suppose *S* is a sub-manifold of *M*. If *S*'s geodesics are always *M*'s geodesics, then we say that the sub-manifold *S* is *totally geodesic*.

If *S* is totally geodesic, then the geodesic distance on *S* is equal to the geodesic distance on *M*. If *S* is not totally geodesic, then the geodesic distance on *S* is an *upper bound* of the geodesic distance on *M*.

Curvature

In *Riemannian Geometry*, *curvature tensor* is defined as the following:

$$R(X,Y,Z,W) = g(\mathcal{R}(Z,W)X,Y) = g((\nabla_Z\nabla_W - \nabla_W\nabla_Z - \nabla_{[Z,W]})X,Y)$$

X,Y,Z,W are vector fields on the manifold which is equipped with the Riemannian metric g. $\mathcal{R}$ is the curvature operator. ∇ is the covariant derivative, and $[Z,W] = Z\circ W - W\circ Z$ is Lie bracket.

Given $X,Y \in T_pM,\ \|X\wedge Y\| \neq 0,\ K(X,Y) = -\frac{R(X,Y,X,Y)}{\|X\wedge Y\|^2}$ is called the *sectional curvature*.

If $K(X,Y) = c$ is a constant for $\forall p \in M, \forall X,Y \in T_pM$, then (M,g) is called the *constant curvature space*.

A *complete*, *simply connected* constant curvature space is *isometrically isomorphic* to (1) the Euclidean Space $\mathbb{R}^n$(*c*=0), (2) *n*-dimensional spheres S^n(*c>0*), or (3) hyperbolic space $H^n(c<0)$. This could facilitate the computation of geodesics for constant curvature spaces.

For example, the statistical manifold for univariate normal distributions is a constant curvature space with $c = -\frac{1}{2}$ [2], which means that it is equipped with the hyperbolic geometry and can be isometrically embedded into the *Poincaré Plane*. This could help us calculate the length of the geodesics without explicitly calculating the connection coefficients Γ_{ij}^k and solving the differential equations.

Note: Fundamentals of Riemannian geometry can be found in any introductory textbooks such as [3].

Relations with f-divergences: Asymptotic Approximation

It is shown that Fisher's information distance (D_F) could be asymptotically approximated by certain functions of KL divergence (D_{KL}) or Hellinger's distance (D_H).

For KL divergence, we have:

$$D_{KL}(p||q) = \frac{1}{2}D_F^2(p||q) + O(D_F^3(p||q)) \text{ [4]}$$

which means that

$$\sqrt{2D_{KL}(p||q)} \to D_F(p||q)$$

as $p \to q$ [5].

Similarly, for Hellinger's distance, we have:

$$D_H(p||q) = 2\sin\left(\frac{D_F(p||q)}{4}\right) = \frac{1}{2}D_F(p||q) + O(D_F^3(p||q)) \text{ [4]}$$

which means that

$$2D_H(p||q) \to D_F(p||q)$$

as $p \to q$ [5].

Some of the researchers have made numerical comparisons between D_F and D_{KL} [6].

Some of the researchers have proposed to approximately calculate the length of the geodesics on the statistical manifolds by *asymptotic approximations* [5][7]:

To approximately calculate the distance between two distributions, $P(x)$ and $Q(x)$, one can densely sample on the statistical manifolds and obtain a family of distributions, $\{p_i(x)\}_{i=1}^N$. For each sample $p_i(x)$, one can approximately calculate its Fisher information distance to the nearby samples $p_j(x)$ by using KL-divergence. Hence, we obtain a weighted graph connecting $P(x), Q(x)$ and $\{p_i(x)\}_{i=1}^N$, and we could obtain the distance between $P(x)$ and $Q(x)$ by finding the shortest path on the graph by Dijkstra's algorithm.

2. Formulation of the Problem

Suppose X, Y are n dimensional Gaussians with $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Our objective is to find $A \in \mathbb{R}^{r \times n}$ so that:

$$A = \arg\max_{A \in \mathbb{R}^{r \times n}} D_F(AX, AY)$$

This objective prompts us to find closed-form formulas for the distance $D_F(P, Q)$ given arbitrary r dimensional Gaussians, P and Q.

Depending on (1) whether $\mu_1 = \mu_2$ or not, and (2) whether $r = 1$ or $r > 1$, in the following we divide our discussion into 4 parts:

(1) Zero-mean and one dimensional cases with $\mu_1 = \mu_2$ and $r = 1$;

(2) Zero-mean and high dimensional cases with $\mu_1 = \mu_2$ and $r > 1$;

(3) Non-zero-mean and one dimensional cases with $\mu_1 \neq \mu_2$ and $r = 1$;

(4) Non-zero-mean and high dimensional cases with $\mu_1 \neq \mu_2$ and $r > 1$.

3. Zero-Mean and One Dimensional Cases

A Straight-forward Approach for Obtaining Geodesics and Fisher Information Distance

To calculate the geodesics, a straight-forward approach would be:

(1) To calculate the Riemannian metric g by the formula:

$$g_{ij} = g\left(\frac{\partial}{\partial \xi_i}, \frac{\partial}{\partial \xi_j}\right) = I(\xi)_{ij} = \mathbb{E}_\xi\left[\frac{\partial}{\partial \xi_i}\log(p(x;\xi))\frac{\partial}{\partial \xi_j}\log(p(x;\xi))\right] \quad ①$$

(2) Given Riemannian metric g, calculate the connection coefficients Γ_{ij}^k by:

$$\Gamma_{ij}^k = \frac{1}{2}\sum_l g^{kl}\left(\frac{\partial g_{il}}{\partial \xi_j} + \frac{\partial g_{lj}}{\partial \xi_i} - \frac{\partial g_{ij}}{\partial \xi_l}\right) \quad ②$$

(3) Given Γ_{ij}^k, solve the following differential equations to obtain the geodesics $\gamma(\mathrm{t})$:

$$\frac{d^2\xi_k(t)}{dt^2} + \sum_{i,j}\Gamma_{ji}^k(\gamma(t))\frac{d\xi_i(t)}{dt}\frac{d\xi_j(t)}{dt} = 0, 1 \leq k \leq n. \quad ③$$

(4) Given the Riemannian metric g and the geodesic $\gamma(t)$, we could finally obtain the length of the geodesics by:

$$D_F(\gamma(t_1)||\gamma(t_2)) = \int_{t_1}^{t_2}\left[\left(\gamma'(t)\right)^T G\left(\gamma(t)\right)\gamma'(t)\right]^{\frac{1}{2}} dt \quad ④$$

The above-mentioned approach could be applied to the discussion of zero-mean and one dimensional cases.

Fisher Information Distance for Zero-Mean and One Dimensional Cases

Univariate normal distributions $\mathcal{N}(\mu, \sigma^2)$ form a two-dimensional statistical manifold.

Suppose that there is a curve $\gamma(\mathrm{t}) = \mathcal{N}(\mu(t) \quad \sigma^2(t))$ on the statistical manifold. It is easy to verify that Riemannian metric $g(t) = \begin{pmatrix} \frac{1}{\sigma^2(t)} & 0 \\ 0 & \frac{2}{\sigma^2(t)} \end{pmatrix}$ by ①:

$$\log p(x) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \log p(x)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

$$\frac{\partial \log p(x)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$$

$$g_{11} = E\left[\left(\frac{\partial \log p(x)}{\partial \mu}\right)^2\right] = \frac{1}{\sigma^4}E[(x-\mu)^2] = \frac{1}{\sigma^2}$$

$$g_{22} = E\left[\left(\frac{\partial \log p(x)}{\partial \sigma}\right)^2\right] = \frac{1}{\sigma^6}E[(x-\mu)^4] - \frac{2}{\sigma^4}E[(x-\mu)^2] + \frac{1}{\sigma^2} = \frac{2}{\sigma^2}$$

$$g_{12} = g_{21} = E\left[\frac{\partial \log p(x)}{\partial \mu}\frac{\partial \log p(x)}{\partial \sigma}\right] = -\frac{1}{\sigma^3}E[x-\mu] + \frac{1}{\sigma^5}E[(x-\mu)^3] = 0$$

For $G^{-1} = (g^{ij})$, we have $g^{11} = \sigma^2, g^{22} = \frac{\sigma^2}{2}$, and $g^{12} = g^{21} = 0$.

Given the Riemannian metric $g(t)$, we could obtain the connection coefficients $\Gamma_{ij}^k(t)$ by ②:

$$\Gamma_{11}^1 = \frac{1}{2}g^{11}\frac{\partial g_{11}}{\partial \mu} = 0$$

$$\Gamma_{11}^2 = -\frac{1}{2}g^{22}\frac{\partial g_{11}}{\partial \sigma} = \frac{1}{2\sigma}$$

$$\Gamma_{12}^1 = \Gamma_{21}^1 = \frac{1}{2}g^{11}\frac{\partial g_{11}}{\partial \sigma} = -\frac{1}{\sigma}$$

$$\Gamma_{12}^2 = \Gamma_{21}^2 = \frac{1}{2}g^{22}\frac{\partial g_{22}}{\partial \mu} = 0$$

$$\Gamma_{22}^1 = -\frac{1}{2}g^{11}\frac{\partial g_{22}}{\partial \mu} = 0$$

$$\Gamma_{22}^2 = \frac{1}{2}g^{22}\frac{\partial g_{22}}{\partial \sigma} = -\frac{1}{\sigma}$$

Suppose $\gamma(\mathrm{t}) = \mathcal{N}(\mu(t) \quad \sigma^2(t))$ is the geodesic connecting $\gamma(0) = \mathcal{N}(\mu_0 \quad \sigma_0^2)$ and $\gamma(1) = \mathcal{N}(\mu_0 \quad \sigma_1^2)$, then by ③ we obtain the following differential equations:

$$\begin{cases} \mu''(t) - \frac{2}{\sigma}\mu'(t)\sigma'(t) = 0 \\ \sigma''(t) + \frac{1}{2\sigma}\left(\mu'(t)\right)^2 - \frac{1}{\sigma}(\sigma'(t))^2 = 0 \end{cases}$$

Assume that $\mu'(t) = 0$, we have:

$$\sigma''(t) - \frac{1}{\sigma}(\sigma'(t))^2 = 0$$

which is equivalent to:

$$\frac{\sigma''(t)}{\sigma'(t)} = \frac{\sigma'(t)}{\sigma(t)}$$

Therefore, we have:

$$\mathrm{d}\big(\log\sigma'(t)\big) = \mathrm{d}(\log\sigma(t))$$

which means that

$$\log\sigma'(t) = \log\sigma(t) + \mathrm{C}$$

So, we have

$$\sigma'(\mathrm{t}) = \mathrm{C}\cdot\sigma(t)$$

which indicates that

$$\mathrm{d}\big(\log\sigma(t)\big) = \frac{\sigma'(t)}{\sigma(t)} = \mathrm{C}$$

So finally we have

$$\log\sigma(t) = C_1 t + C_2$$

and

$$\sigma(t) = C_1 e^{C_2 t}$$

In conclusion, $\mu(t) = \mu_0$ and $\sigma(t) = \sigma_0 e^{\log\frac{\sigma_1}{\sigma_0}t} = \sigma_0\left(\frac{\sigma_1}{\sigma_0}\right)^t$ is the desired solution. $\gamma(t) = \mathcal{N}(\mu_0 \quad \sigma_0^2\left(\frac{\sigma_1}{\sigma_0}\right)^{2t})$ is the geodesic connecting $\mathcal{N}(\mu_0 \quad \sigma_0^2)$ and $\mathcal{N}(\mu_0 \quad \sigma_1^2)$, and by ④ we could obtain the length of the geodesic:

$$D_F\big(\mathcal{N}(\mu_0 \quad \sigma_0^2)||\mathcal{N}(\mu_0 \quad \sigma_1^2)\big) = L\big(\gamma(t)\big) = \int_0^1 [g_{22}(t)(\sigma'(t))^2]^{\frac{1}{2}}dt = \sqrt{2}\left|\log\frac{\sigma_1}{\sigma_0}\right|$$

$$= \sqrt{2}\log\frac{\max\{\sigma_0,\sigma_1\}}{\min\{\sigma_0,\sigma_1\}} = \frac{1}{\sqrt{2}}\log\frac{\max\{\sigma_0^2,\sigma_1^2\}}{\min\{\sigma_0^2,\sigma_1^2\}}$$

Linear Dimensionality Reduction with $\boldsymbol{\mu_1 = \mu_2}$ and $\boldsymbol{r = 1}$

Suppose $X\sim\mathcal{N}(0, I_n)$ and $Y\sim\mathcal{N}(0,\Sigma)$, our objective is to find an optimal $A \in \mathbb{R}^{1\times n}$ so that $D_F(AX\|AY)$ is maximized.

From the above-mentioned formula we obtain that

$$D_F(AX\|AY) = \frac{1}{\sqrt{2}}\log\frac{\max\{AA^T, A\Sigma A^T\}}{\min\{AA^T, A\Sigma A^T\}}$$

It is easy to verify that $D_F(AX\|AY)$ will be invariant when the matrix A changes by a scalar factor. Henceforth, we could always assume that $AA^T = 1$ without any loss of generality.

Therefore, our objective becomes finding *A* that satisfies the following:

$$A = \arg\max_{AA^T=1}\frac{\max\{1, A\Sigma A^T\}}{\min\{1, A\Sigma A^T\}}$$

The solution is exactly the same to the solution to maximizing Hellinger distance.

4. Zero-Mean and High Dimensional Cases

Parametrizations for Multivariate Gaussian Distributions

For multivariate normal distributions, we have different ways of parametrizations. Taking the symmetry condition $\sigma_{ij} = \sigma_{ji}$ into consideration, we could parametrize *n* dimensional normal distributions as $(\dots,\mu_i,\dots,\dots,\sigma_{pq},\dots)(1\le i\le n, 1\le p\le q\le n)$ on a $n + \frac{n(n+1)}{2}$ dimensional manifold. [2] has made a comprehensive discussion about the geometry of these statistical manifolds.

An alternative approach parametrizes *n* dimensional normal distributions as

$(\dots, \mu_i, \dots, \dots, \sigma_{pq}, \dots)(1 \le i \le n, 1 \le p, q \le n)$ on a $n + n^2$ dimensional manifold. For this section, we use the second method of parametrization for our calculations.

Riemannian Metrics and Connection Coefficients for Zero-Mean and High Dimensional Cases

First, we calculate the Riemannian metric g:

$$\log p(x) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$\frac{\partial \log p(x)}{\partial \mu_i} = \sum_{j=1}^{n} \sigma^{ij}(x_j - \mu_j)$$

$$\frac{\partial \log p(x)}{\partial \sigma_{ij}} = -\frac{1}{2}\sigma^{ij} + \frac{1}{2}\sum_{k,l} \sigma^{ik}(x_k - \mu_k)(x_l - \mu_l)\sigma^{lj}$$

$$g_{\mu_i\mu_j} = \mathrm{E}\left[\frac{\partial \log p(x)}{\partial \mu_i}\frac{\partial \log p(x)}{\partial \mu_j}\right] = \sum_{k,l} \sigma^{ik} E[(x_k - \mu_k)(x_l - \mu_l)]\sigma^{lj} = (\Sigma^{-1}\Sigma\Sigma^{-1})_{ij} = \sigma^{ij}$$

$$\begin{aligned} g_{\sigma_{ij}\sigma_{kl}} &= \mathrm{E}\left[\frac{\partial \log p(x)}{\partial \sigma_{ij}}\frac{\partial \log p(x)}{\partial \sigma_{kl}}\right] \\ &= \frac{1}{4}\sigma^{ij}\sigma^{kl} - \frac{1}{4}\sum_{r,s} \sigma^{ij}\sigma^{kr} E[(x_r - \mu_r)(x_s - \mu_s)]\sigma^{sl} \\ &\quad - \frac{1}{4}\sum_{p,q} \sigma^{kl}\sigma^{ip} E\big[(x_p - \mu_p)(x_q - \mu_q)\big]\sigma^{qj} \\ &\quad + \frac{1}{4}\sum_{p,q,r,s} \sigma^{ip}\sigma^{qj}\sigma^{kr}\sigma^{sl} E\big[(x_p - \mu_p)(x_q - \mu_q)(x_r - \mu_r)(x_s - \mu_s)\big] \\ &= -\frac{1}{4}\sigma^{ij}\sigma^{kl} \\ &\quad + \frac{1}{4}\sum_{p,q,r,s} \sigma^{ip}\sigma^{qj}\sigma^{kr}\sigma^{sl} E\big[(x_p - \mu_p)(x_q - \mu_q)(x_r - \mu_r)(x_s - \mu_s)\big] \\ &= \frac{1}{4}\sigma^{ik}\sigma^{jl} + \frac{1}{4}\sigma^{il}\sigma^{jk} \end{aligned}$$

$$\begin{aligned} g_{\mu_i\sigma_{jk}} &= \mathrm{E}\left[\frac{\partial \log p(x)}{\partial \mu_i}\frac{\partial \log p(x)}{\partial \sigma_{jk}}\right] \\ &= -\frac{1}{2}\sum_{l} \sigma^{il}\sigma^{jk} E[x_l - \mu_l] \\ &\quad + \frac{1}{2}\sum_{p,q,l} \sigma^{il}\sigma^{jp}\sigma^{qk} E\big[(x_l - \mu_l)(x_p - \mu_p)(x_q - \mu_q)\big] = 0 \end{aligned}$$

Then we calculate and simplify the expressions for the connection coefficients Γ_{ij}^{k}:

$$\Gamma_{\mu_i\mu_j}^{\mu_k} = \frac{1}{2}\sum_{l} g^{\mu_k\mu_l}\left(\frac{\partial g_{\mu_i\mu_l}}{\partial \mu_j} + \frac{\partial g_{\mu_l\mu_j}}{\partial \mu_i} - \frac{\partial g_{\mu_i\mu_j}}{\partial \mu_l}\right) = 0$$

$$\Gamma_{\mu_i\mu_j}^{\sigma_{kl}} = \frac{1}{2}\sum_{m,n} g^{\sigma_{kl}\sigma_{mn}}\left(\frac{\partial g_{\mu_i\sigma_{mn}}}{\partial \mu_j} + \frac{\partial g_{\sigma_{mn}\mu_j}}{\partial \mu_i} - \frac{\partial g_{\mu_i\mu_j}}{\partial \sigma_{mn}}\right) = -\frac{1}{2}\sum_{m,n} g^{\sigma_{kl}\sigma_{mn}}\frac{\partial g_{\mu_i\mu_j}}{\partial \sigma_{mn}}$$

$$\Gamma_{\mu_i\sigma_{kl}}^{\mu_j} = \frac{1}{2}\sum_{m} g^{\mu_j\mu_m}\left(\frac{\partial g_{\mu_i\mu_m}}{\partial \sigma_{kl}} + \frac{\partial g_{\mu_m\sigma_{kl}}}{\partial \mu_i} - \frac{\partial g_{\mu_i\sigma_{kl}}}{\partial \mu_m}\right) = \frac{1}{2}\sum_{m} g^{\mu_j\mu_m}\frac{\partial g_{\mu_i\mu_m}}{\partial \sigma_{kl}}$$

$$\Gamma_{\mu_i\sigma_{kl}}^{\sigma_{mn}} = \frac{1}{2}\sum_{p,q} g^{\sigma_{mn}\sigma_{pq}}\left(\frac{\partial g_{\mu_i\sigma_{pq}}}{\partial \sigma_{kl}} + \frac{\partial g_{\sigma_{pq}\sigma_{kl}}}{\partial \mu_i} - \frac{\partial g_{\mu_i\sigma_{kl}}}{\partial \sigma_{pq}}\right) = 0$$

$$\Gamma_{\sigma_{ij}\sigma_{kl}}^{\mu_m} = \frac{1}{2}\sum_{n} g^{\mu_m\mu_n}\left(\frac{\partial g_{\sigma_{ij}\mu_m}}{\partial \sigma_{kl}} + \frac{\partial g_{\mu_m\sigma_{kl}}}{\partial \sigma_{ij}} - \frac{\partial g_{\sigma_{ij}\sigma_{kl}}}{\partial \mu_m}\right) = 0$$

$$\Gamma_{\sigma_{ij}\sigma_{kl}}^{\sigma_{mn}} = \frac{1}{2}\sum_{p,q} g^{\sigma_{mn}\sigma_{pq}}\left(\frac{\partial g_{\sigma_{ij}\sigma_{pq}}}{\partial \sigma_{kl}} + \frac{\partial g_{\sigma_{pq}\sigma_{kl}}}{\partial \sigma_{ij}} - \frac{\partial g_{\sigma_{ij}\sigma_{kl}}}{\partial \sigma_{pq}}\right)$$

Then we could obtain the differential equations for the geodesics:

$$\begin{cases} \dfrac{d^2\mu_k(t)}{dt^2} + \displaystyle\sum_{j,p,q} \Gamma_{\mu_j\sigma_{pq}}^{\mu_k}\frac{d\mu_j(t)}{dt}\frac{d\sigma_{pq}(t)}{dt} = 0 \\ \dfrac{d^2\sigma_{mn}(t)}{dt^2} + \displaystyle\sum_{i,j} \Gamma_{\mu_i\mu_j}^{\sigma_{mn}}\frac{d\mu_i(t)}{dt}\frac{d\mu_j(t)}{dt} + \sum_{i,j,k,l} \Gamma_{\sigma_{ij}\sigma_{kl}}^{\sigma_{mn}}\frac{d\sigma_{ij}(t)}{dt}\frac{d\sigma_{kl}(t)}{dt} = 0 \end{cases}$$

The Solutions to the Geodesics in Certain Circumstances

It seems that the above-mentioned differential equations are far from tractable, however, it turns out that we could obtain closed-form solutions for the geodesics connecting $X \sim \mathcal{N}(\mu, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu, \Sigma_2)$ if Σ_1 and Σ_2 are two diagonal matrices.

All we need is to tentatively assume that $\forall i, \mu_i'(t) = 0$ and $\forall i \neq j, \sigma_{ij}'(t) = 0$ first, and then try to solve the differential equations. It turns out that the differential equations have solutions that satisfy our assumptions, which in turn proves that our assumptions are correct.

Given $\forall i, \mu_i'(t) = 0$ and $\forall i \neq j, \sigma_{ij}'(t) = 0$, the above-mentioned differential equations become:

$$\frac{d^2\sigma_{mn}(t)}{dt^2} + \sum_{i,j} \Gamma_{\sigma_{ii}\sigma_{jj}}^{\sigma_{mn}}\frac{d\sigma_{ii}(t)}{dt}\frac{d\sigma_{jj}(t)}{dt} = 0$$

It can be proved that:

$$\Gamma_{\sigma_{ii}\sigma_{jj}}^{\sigma_{mn}} = \begin{cases} -\dfrac{1}{\sigma_{ii}} & i = j = m = n \\ 0 & otherwise \end{cases}$$

So the differential equations become:

$$\begin{cases} \dfrac{d^2\sigma_{mn}(t)}{dt^2} = 0 & m \neq n \\ \dfrac{d^2\sigma_{mm}(t)}{dt^2} - \dfrac{1}{\sigma_{mm}(t)}\left(\dfrac{d\sigma_{mm}(t)}{dt}\right)^2 = 0 & m = n \end{cases}$$

It has been shown in the previous section that the solution to $\frac{d^2\sigma_{mm}(t)}{dt^2} - \frac{1}{\sigma_{mm}(t)}\left(\frac{d\sigma_{mm}(t)}{dt}\right)^2 =$

0 is:

$$\sigma_{mm}(t) = C_{m1}e^{C_{m2}t}$$

So the differential equations indeed have solutions under our assumptions. Therefore, we obtained the solutions to the geodesics connecting $X \sim \mathcal{N}(\mu, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu, \Sigma_2)$ when Σ_1 and Σ_2 are two diagonal matrices. For the geodesics $\gamma(t) = (\dots, \mu_i(t), \dots, \dots, \sigma_{ij}(t), \dots)$, $\mu_i(t)$ and $\sigma_{ij}(t)(i \neq j)$ should be stationary and $\sigma_{ii}(t)$ should take the form of $\sigma_{ii}(t) = C_{i1}e^{C_{i2}t}$. This confirms that the submanifold:

$$\{\mathcal{N}(\mu_0, \Sigma) | \mu_0 \in \mathbb{R}^{n\times 1}, \Sigma \in \mathbb{R}^{n\times n} \ is \ a \ diagonal, positive \ definite \ matrix\}$$

is totally geodesic as is stated in [2].

Suppose that $X \sim \mathcal{N}(\mu, \Lambda_1)$, $Y \sim \mathcal{N}(\mu, \Lambda_2)$, Λ_1, Λ_2 are diagonal matrices that $\Lambda_1 = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ and $\Lambda_2 = \begin{pmatrix} \lambda_1' & & \\ & \ddots & \\ & & \lambda_n' \end{pmatrix}$. Based on previous discussions, the geodesic $\gamma(t) = (\dots, \mu_i(t), \dots, \dots, \sigma_{jk}(t), \dots)$ that connects $\mathcal{N}(\mu, \Lambda_1)$ and $\mathcal{N}(\mu, \Lambda_2)$ should satisfy the following:

$$\mu_i(t) = \mu_i, \forall i$$
$$\sigma_{ij}(t) = 0, \forall i \neq j$$
$$\sigma_{ii}(t) = \lambda_i \left(\frac{\lambda_i'}{\lambda_i}\right)^t, \forall i, \forall t \in [0,1]$$

Hence, we can calculate the length of the geodesic:

$$D_F(\mathcal{N}(\mu, \Lambda_1) \| \mathcal{N}(\mu, \Lambda_2)) = L\big(\gamma(t)\big) = \int_0^1 \left[\sum_{i,j} \sigma_{ii}'(t)\sigma_{jj}'(t) g_{\sigma_{ii}\sigma_{jj}}(t)\right]^{\frac{1}{2}} dt$$

Since:

$$g_{\sigma_{ii}\sigma_{jj}} = \frac{1}{2}\big(\sigma^{ij}\big)^2 = \begin{cases} 0 & i \neq j \\ \dfrac{1}{2\sigma_{ii}^2} & i = j \end{cases}$$

we are able to obtain the closed form solution to Fisher information distance between $\mathcal{N}(\mu, \Lambda_1)$ and $\mathcal{N}(\mu, \Lambda_2)$:

$$D_F(\mathcal{N}(\mu, \Lambda_1) \| \mathcal{N}(\mu, \Lambda_2)) = L\big(\gamma(t)\big) = \int_0^1 \left[\sum_i \big(\sigma_{ii}'(t)\big)^2 g_{\sigma_{ii}\sigma_{ii}}(t)\right]^{\frac{1}{2}} dt = \frac{1}{\sqrt{2}} \sqrt{\sum_i \left(\log(\frac{\lambda_i'}{\lambda_i})\right)^2}$$

Linear Dimensionality Reduction with $\boldsymbol{\mu_1 = \mu_2}$ and $\boldsymbol{r > 1}$

Suppose that $X \sim \mathcal{N}(0, I_n)$, $Y \sim \mathcal{N}(0, \Sigma)$, our objective is to find $A \in \mathbb{R}^{r\times n} (rank(A) = r, 1 < r < n)$ so that $D_F(AX \| AY)$ is maximized.

By *QR decomposition*, we have $A = RU$, where $R \in \mathbb{R}^{r\times n}$ is a semi-orthogonal matrix satisfying $RR^T = I_r$, and $U \in \mathbb{R}^{n\times n}$ is a lower triangular matrix with $\text{rank}(U) = n$.

Here we make an *assumption* that Fisher information distance is invariant under full-rank lower triangular transformations:

$$D_F(X \| Y) = D_F(UX \| UY)$$

This is intuitive, because UX and UY's row vectors are just linear combinations of X and Y's row vectors. Since U is of full rank, this transformation should not produce any loss or gain of information.

If the above-mentioned assumption is correct, then we could constrain ourselves to the cases when $A \in \mathbb{R}^{r\times n}$ is semi-orthogonal without any loss of generality.

Hence, our objective becomes finding semi-orthogonal matrices A that satisfy the following:

$$A = \arg\max_{AA^T=I_r} D_F(AX\|AY)$$

where

$$D_F(AX\|AY) = D_F(\mathcal{N}(0, I_r)\|\mathcal{N}(0, A\Sigma A^T))$$

Suppose Σ's eigenvalues are $\lambda_1, \dots, \lambda_n$, and $A\Sigma A^T$'s eigenvalues are $\gamma_1, \dots, \gamma_r$. Since Fisher information distance is invariant under orthogonal transformations [7], we could obtain that:

$$D_F(\mathcal{N}(0, I_r)\|\mathcal{N}(0, A\Sigma A^T)) = D_F(\mathcal{N}(0, I_r)\|\mathcal{N}(0, \Gamma))$$

where $\Gamma = \begin{pmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_r \end{pmatrix}$ is a diagonal matrix.

Hence we could obtain Fisher information distance $D_F(AX\|AY)$ by using the formula:

$$D_F(AX\|AY) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\log\gamma_i)^2}$$

Similar to discussions with *f*-divergences, by *Poincaré Separation Theorem or Cauchy Interlacing Theorem* we have $\lambda_{n-r+i} \le \gamma_i \le \lambda_i$. To maximize $D_F(AX\|AY)$, we need to individually maximize each $(\log\gamma_i)^2$. This leads to a solution that is exactly the same to the solution to maximizing Hellinger distance when $\mu_1 = \mu_2$ and $r > 1$.

To summarize, in zero-mean cases, linear dimensionality reduction with the objective of maximizing Fisher information distance will produce an outcome that is exactly the same to the one that is produced by linear dimensionality reduction with the objective of maximizing Hellinger distance.

5. Non-Zero-Mean and One Dimensional Cases

Calculations Following the Straight-forward Approach

If we want to obtain the geodesics in non-zero-mean cases following the above-mentioned straight-forward approach, we will obtain the following differential equations:

$$\begin{cases} \dfrac{d^2\mu(t)}{dt^2} - \dfrac{2}{\sigma(t)}\dfrac{d\mu(t)}{dt}\dfrac{d\sigma(t)}{dt} = 0 \\ \dfrac{d^2\sigma(t)}{dt^2} + \dfrac{1}{2\sigma(t)}\left(\dfrac{d\mu(t)}{dt}\right)^2 - \dfrac{1}{\sigma(t)}\left(\dfrac{d\sigma(t)}{dt}\right)^2 = 0 \end{cases}$$

which is equivalent to:

$$\begin{cases} \mu''\sigma - 2\mu'\sigma' = 0 & ⑤ \\ \sigma''\sigma + \dfrac{(\mu')^2}{2} - (\sigma')^2 = 0 & ⑥ \end{cases}$$

From ⑤ we could obtain that:

$$\frac{\mu''}{\mu'} = \frac{2\sigma'}{\sigma} \Rightarrow (\log\mu')' = 2(\log\sigma)' \Rightarrow \log\mu' = 2\log\sigma + \mathrm{C} \Rightarrow \mu' = C\sigma^2$$

Replacing μ' by $C\sigma^2$ in ⑥, we could obtain:

$$\sigma''\sigma + C\sigma^4 - (\sigma')^2 = 0$$

which is equivalent to:

$$\frac{\sigma''}{\sigma} + C\sigma^2 - \left(\frac{\sigma'}{\sigma}\right)^2 = 0$$

Noticing that $\frac{\sigma''}{\sigma} - \left(\frac{\sigma'}{\sigma}\right)^2 = (\frac{\sigma'}{\sigma})'$, we have:

$$\left(\frac{\sigma'}{\sigma}\right)' = C\sigma^2$$

Multiplying the above equation by $\frac{\sigma'}{\sigma}$ on each side, we obtain:

$$\left(\frac{\sigma'}{\sigma}\right)' \frac{\sigma'}{\sigma} = C\sigma\sigma'$$

Noticing that $\mathrm{d}u^2 = 2uu'$, we have:

$$d\left[\left(\frac{\sigma'}{\sigma}\right)^2\right] = Cd(\sigma^2)$$

Hence we have:

$$\left(\frac{\sigma'}{\sigma}\right)^2 = C_1\sigma^2 + C_2$$

So

$$\frac{d\sigma}{dt} = \sigma' = \sigma\sqrt{C_1\sigma^2 + C_2}$$

Therefore, we have

$$\frac{d\sigma}{\sigma\sqrt{C_1\sigma^2 + C_2}} = dt$$

or

$$\int \frac{d\sigma}{\sigma\sqrt{C_1\sigma^2 + C_2}} = t + C_3$$

This integral is in fact tractable. In principle, we could obtain $\sigma(t)$ and $\mu(t)$ following this approach, however, the calculation is sophisticated and it is difficult for us to calculate the length of the geodesics. Due to the difficulties in calculations, we resort to an alternative approach which could simplify the calculations.

Discussions in an Alternative Perspective

[2] has pointed out that the statistical manifold for univariate normal distributions is a constant curvature space with constant negative sectional curvature $c = -\frac{1}{2}$. Since the statistical manifold for univariate normal distributions is complete, simply connected, with constant negative sectional curvature, it is equipped with the hyperbolic geometry which could simplify our calculations.

[6] has made full use of the hyperbolic geometry to calculate the geodesic distance on the statistical manifolds. Recognizing the resemblance between the metric of Poincaré plane and the metric of the statistical manifold for univariate normal distributions, the authors of [6] were able to obtain the closed form formula for the geodesic distance between $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$:

$$D_F(\mathcal{N}(\mu_1, \sigma_1^2) \| \mathcal{N}(\mu_2, \sigma_2^2)) = \sqrt{2}\log\frac{\left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right)\right\|_2 + \left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right\|_2}{\left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, -\sigma_2\right)\right\|_2 - \left\|\left(\frac{\mu_1}{\sqrt{2}}, \sigma_1\right) - \left(\frac{\mu_2}{\sqrt{2}}, \sigma_2\right)\right\|_2}$$

where $\|\cdot\|_2$ is the L^2-norm.

Linear Dimensionality Reduction with $\boldsymbol{\mu_1 \neq \mu_2}$ and $\boldsymbol{r = 1}$: the Objective

Suppose $X \sim \mathcal{N}(0, I_n)$ and $Y \sim \mathcal{N}(\mu, \Sigma)$. Our aim is to find an optimal $A \in \mathbb{R}^{1\times n}$ so that $D_F(AX||AY)$ achieves its maximal possible value.

It is easy to verify that $D_F(AX\|AY)$ is invariant if A changes by a scalar factor, so we could always assume that $AA^T = 1$ without loss of generality.

Hence, our objective is to find $A \in \mathbb{R}^{1\times n}$ that satisfies the following:

$$A = \arg\max_{AA^T=1} D_F(AX\|AY)$$

Now let's take a closer look at the formula for $D_F(\mathcal{N}(\mu_1,\sigma_1^2)\|\mathcal{N}(\mu_2,\sigma_2^2))$. Suppose $AX\sim\mathcal{N}(\mu_1,\sigma_1^2)$ and $AY\sim\mathcal{N}(\mu_2,\sigma_2^2)$, and for simplicity we denote $\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\|_2$ by d_1 and $\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|_2$ by d_2, then we have:

$$D_F(AX\|AY) = \sqrt{2}\log\frac{d_1+d_2}{d_1-d_2}$$

To maximize $D_F(AX\|AY)$, it is equivalent to maximizing $\frac{d_1+d_2}{d_1-d_2}$. Since $\frac{d_1+d_2}{d_1-d_2} = 1+\frac{2d_2}{d_1-d_2}$, it is equivalent to maximizing $\frac{d_2}{d_1-d_2}$.

Since it is guaranteed that $d_1 - d_2 > 0$, we could equivalently seek to minimize $\frac{d_1-d_2}{d_2}$, which is equivalent to minimizing $\frac{d_1}{d_2}$ or minimizing $\left(\frac{d_1}{d_2}\right)^2$.

We have:

$$\left(\frac{d_1}{d_2}\right)^2 = \frac{\frac{(\mu_1-\mu_2)^2}{2}+(\sigma_1+\sigma_2)^2}{\frac{(\mu_1-\mu_2)^2}{2}+(\sigma_1-\sigma_2)^2} = 1+4\cdot\frac{\sigma_1\sigma_2}{\frac{(\mu_1-\mu_2)^2}{2}+(\sigma_1-\sigma_2)^2}$$

So, we could equivalently seek to minimize $\frac{\sigma_1\sigma_2}{\frac{(\mu_1-\mu_2)^2}{2}+(\sigma_1-\sigma_2)^2}$, which is equivalent to maximizing the following:

$$\frac{\frac{(\mu_1-\mu_2)^2}{2}+(\sigma_1-\sigma_2)^2}{\sigma_1\sigma_2} = \frac{(\mu_1-\mu_2)^2}{2\sigma_1\sigma_2}+\frac{\sigma_1}{\sigma_2}+\frac{\sigma_2}{\sigma_1}-2$$

So our objective could be equivalently transformed to maximizing the following:

$$\frac{(\mu_1-\mu_2)^2}{2\sigma_1\sigma_2}+\frac{\sigma_1}{\sigma_2}+\frac{\sigma_2}{\sigma_1}$$

Since $X\sim\mathcal{N}(0,I_n)$, $Y\sim\mathcal{N}(\mu,\Sigma)$ and $AA^T=1$, we have $\mu_1=0$, $\mu_2=A\mu$, $\sigma_1^2=AA^T=1$ and $\sigma_2^2=A\Sigma A^T$.

Hence, our objective becomes finding A that satisfies the following:

$$A = \arg\max_{AA^T=1}\left[\sqrt{A\Sigma A^T}+\frac{1+\frac{A\mu\mu^TA^T}{2}}{\sqrt{A\Sigma A^T}}\right]$$

Linear Dimensionality Reduction with $\mu_1 \neq \mu_2$ and $r=1$: the Structure of the Solutions

To reveal certain structure of the solutions, we need to make several assumptions first.

Assume that $\Sigma = V\Lambda V^T$, where $\Lambda = \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}$ is a diagonal matrix with $\lambda_1,\dots,\lambda_n$

being the eigenvalues of Σ, and $V = (v_1 \quad \cdots \quad v_n)$ is an orthogonal matrix with $v_1, \dots, v_n \in \mathbb{R}^{n\times 1}$ being the eigenvectors of Σ.

Assume that $\mu = V\alpha = (v_1 \quad \cdots \quad v_n)\begin{pmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{pmatrix} = \sum_i \alpha_i v_i$ with $\alpha_1, \dots, \alpha_n$ being scalars.

Assume that $A = \beta^T V^{\mathrm{T}} = (\beta_1 \quad \cdots \quad \beta_n)\begin{pmatrix}v_1^T \\ \vdots \\ v_n^T\end{pmatrix} = \sum_i \beta_i v_i^T$ with $\beta_1, \dots, \beta_n$ being scalars.

Since V is orthogonal, the constraint $AA^T = 1$ is equivalent to the constraint of $\beta^{\mathrm{T}}\beta = 1$.

Hence, our previous objective:

$$A = \arg\max_{AA^T=1}\left[\sqrt{A\Sigma A^T} + \frac{1 + \frac{\mu^T A^T A\mu}{2}}{\sqrt{A\Sigma A^T}}\right]$$

can be equivalently transformed to the following objective:

$$\beta = \arg\max_{\|\beta\|_2^2=\beta^T\beta=1}\left[\sqrt{\beta^T\Lambda\beta} + \frac{1 + \frac{\alpha^T\beta\beta^T\alpha}{2}}{\sqrt{\beta^T\Lambda\beta}}\right] \quad ⑦$$

Note that $\|\mu\|_2^2 = \mu^T\mu = \alpha^T V^T V\alpha = \alpha^T\alpha = \|\alpha\|_2^2$. In addition, under the constraint of $\|\beta\|_2^2 = \beta^T\beta = 1$, we have $\alpha^T\beta\beta^T\alpha =< \alpha, \beta >^2 = \|\alpha\|_2^2\|\beta\|_2^2\cos^2\theta = \|\mu\|_2^2\cos^2\theta$, where θ is the angle between the vector of α and the vector of β.

Now consider a family of constrained optimization problems:

$$\beta(\theta) = \arg\max_{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}}\left[\sqrt{\beta^T\Lambda\beta} + \frac{1 + \frac{\|\mu\|_2^2\cos^2\theta}{2}}{\sqrt{\beta^T\Lambda\beta}}\right] \quad ⑧$$

If β^* is the optimal solution to ⑦, then it is necessary that for some θ, β^* is the optimal solution to ⑧.

Note that under the constraint of ⑧, $1 + \frac{\|\mu\|_2^2\cos^2\theta}{2}$ is a constant positive number. For any constant positive number C, the function $f(x) = x + \frac{C}{x}(x > 0)$ is convex, which means that $f(x)$ will reach its maximum only when x reaches the boundaries of the constraints. This means that in order to obtain the optimal solution to ⑧, it is necessary that $\sqrt{\beta^T\Lambda\beta}$ (or equivalently, $\beta^T\Lambda\beta$) reach either the maximum or minimum under the constraints of ⑧.

So, if for some θ, β^* is the optimal solution to ⑧, it is necessary that for the same θ, β^* is the optimal solution to either of the following two problems:

$$\beta^+(\theta) = \arg\max_{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}} \beta^T\Lambda\beta \quad ⑨$$

$$\beta^-(\theta) = \arg\min_{\substack{\|\beta\|_2^2=\beta^T\beta=1 \\ <\frac{\alpha}{\|\alpha\|_2},\beta>=cos\theta}} \beta^T\Lambda\beta \quad ⑩$$

In order to obtain the solutions to ⑨ and ⑩, we just need the Lagrange Multiplier Method.

The Lagrange function takes the following form:

$$L(\beta, l_1, l_2) = \beta^{\mathrm{T}} \Lambda \beta + l_1(\beta^T \beta - 1) + l_2(\tilde{\alpha}^T \beta - cos\theta)$$

where $\tilde{\alpha} = \frac{\alpha}{\|\alpha\|_2}$.

If β^* is the optimal solution to ⑨ or ⑩, it is necessary that $\frac{\partial L(\beta, l_1, l_2)}{\partial \beta} = 0$:

$$\frac{\partial L(\beta, l_1, l_2)}{\partial \beta} = 2\Lambda\beta + 2l_1\beta + l_2\tilde{\alpha} = 0$$

So we have:

$$(\Lambda + l_1 I_n)\beta = -\frac{l_2}{2\|\alpha\|_2}\alpha$$

which means that for some parameters γ, k, we have:

$$(\lambda_i + \gamma)\beta_i = k\alpha_i (1 \le i \le n)$$

Hence, we find out that in $\beta = (\beta_1, \dots, \beta_n)^{\mathrm{T}}$, $\beta_1, \dots, \beta_n$ are proportional to each other:

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

Since $\beta^T \beta = 1$, β will be uniquely determined if we could decide γ.

However, problem ⑨ and ⑩ are only *relaxations* of problem ⑦, and it is not likely that we will be able to decide which γ is for the solution of ⑦ and find its closed forms. What we have discovered is that the solutions to ⑦ is structured: the components of β are proportional to each other, and they are determined by α (which reflects the influence of the difference in means), Λ (which reflects the influence of the difference in covariance matrices), and an undecided parameter, γ.

Linear Dimensionality Reduction with $\boldsymbol{\mu_1 \neq \mu_2}$ and $\boldsymbol{r = 1}$: an Algorithm of the First Order

Our objective is:

$$\max F(A) = \sqrt{A\Sigma A^T} + \frac{1 + \frac{\mu^T A^T A \mu}{2}}{\sqrt{A\Sigma A^T}}$$

$$\text{s.t. } AA^T = 1$$

In Euclidean space, $F(A)$'s gradient is:

$$\nabla_A F = \frac{A(\Sigma + \mu\mu^T)}{(A\Sigma A^T)^{\frac{1}{2}}} - \frac{\left(1 + \frac{\mu^T A^T A \mu}{2}\right) A\Sigma}{(A\Sigma A^T)^{\frac{3}{2}}}$$

Note that on the manifold of $AA^T = 1$, we are allowed to move only along the directions that are orthogonal to the direction of A, so $F(A)$'s gradient on the manifold of $AA^T = 1$ should be:

$$\nabla_A^M F = \nabla_A F - < \nabla_A F, A > A$$

Since we have the constraint $AA^T = 1$, in gradient ascent we cannot simply update the points by $A_{k+1} = A_k + \alpha \nabla_{A_k}^M F$, or we will be violating the constraint. Instead we need the operation of "*retraction*" [8]:

$$A_k^* = A_k + \alpha \nabla_{A_k}^M F$$

$$A_{k+1} = \frac{A_k^*}{\|A_k^*\|_2}$$

which could guarantee that our gradient ascent algorithm will not violate the constraints.

Gradient Ascent Algorithm for One Dimensional Cases

Input: μ, Σ

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

1. Initialize A_0 randomly with the constraint $A_0 A_0^T = 1$; $k \leftarrow 0$;

2. Do:

(1) Compute $\nabla_{A_k} F$;

(2) Compute the gradient on the manifold: $\nabla_{A_k}^M F \leftarrow \nabla_{A_k} F - < \nabla_{A_k} F, A_k > A_k$;

(3) Update A_k in an usual way and obtain an intermediate result: $A_k^* \leftarrow A_k + \alpha \nabla_{A_k}^M F$;

(4) Retraction: $A_{k+1} \leftarrow \frac{A_k^*}{\|A_k^*\|_2}$

(5) $k \leftarrow k + 1$;

While not convergence;

3. Return $A = A_k$.

Connection between Solutions to Zero-Mean Cases and Non-Zero-Mean Cases

Recall that:

$$(\Lambda + l_1 I_n)\beta = -\frac{l_2}{2\|\alpha\|_2}\alpha$$

Since $\Lambda = V^T \Sigma V$, $I_n = V^T V$, $\alpha = V^T \mu$, $\|\alpha\|_2 = \|\mu\|_2$, we have:

$$(V^T \Sigma V + l_1 V^T V)\beta = -\frac{l_2}{2\|\mu\|_2} V^T \mu$$

Since $V\beta = A^T$, we have:

$$(V^T \Sigma + l_1 V^T) A^T = -\frac{l_2}{2\|\mu\|_2} V^T \mu$$

Hence

$$(\Sigma + l_1 I_n) A^T = -\frac{l_2}{2\|\mu\|_2} \mu$$

This is what the optimal solution A should satisfy in the non-zero-mean cases.

In zero-mean cases, we have $\mu = 0$, so the optimal solution A should satisfy:

$$(\Sigma + l_1 I_n) A^T = 0$$

which is the eigenvalue problem. This is consistent with our previous discussion that in zero-mean cases, the solution is exactly the same to the solution to Hellinger distance.

6. Non-Zero-Mean and High Dimensional Cases

An Upper Bound of the Geodesic Distance

To the best of my knowledge, there is no generic closed-form solution for non-zero-mean and high dimensional cases, though partial solutions pertaining to some special cases might exist.

[6] claimed that they could obtain the geodesic distance between $\mathcal{N}(\mu_1, \Lambda_1)$ and $\mathcal{N}(\mu_2, \Lambda_2)$, if Λ_1 and Λ_2 are diagonal matrices. Their claim was based on the observation that the submanifold $S = \{\mathcal{N}(\mu, \Sigma) | \mu \in \mathbb{R}^{n\times 1}, \Sigma \in \mathbb{R}^{n\times n} \ is \ a \ diagonal, positive \ definite \ matrix\}$ is equipped with a metric that is similar to the metric of hyperbolic space H^{2n}.

[6] has obtained the following formula, which is in fact the geodesic distance between $\mathcal{N}(\mu_1, \Lambda_1)$ and $\mathcal{N}(\mu_2, \Lambda_2)$ (Λ_1 and Λ_2 are diagonal matrices) *on the submanifold S*:

$$\widetilde{D_F}(\mathcal{N}(\mu_1,\Lambda_1)\|\mathcal{N}(\mu_2,\Lambda_2))$$
$$=\sqrt{2}\sqrt{\sum_i\left(\log\frac{\left\|\left(\frac{\mu_{1i}}{\sqrt{2}},\sigma_{1i}\right)-\left(\frac{\mu_{2i}}{\sqrt{2}},-\sigma_{2i}\right)\right\|_2+\left\|\left(\frac{\mu_{1i}}{\sqrt{2}},\sigma_{1i}\right)-\left(\frac{\mu_{2i}}{\sqrt{2}},\sigma_{2i}\right)\right\|_2}{\left\|\left(\frac{\mu_{1i}}{\sqrt{2}},\sigma_{1i}\right)-\left(\frac{\mu_{2i}}{\sqrt{2}},-\sigma_{2i}\right)\right\|_2-\left\|\left(\frac{\mu_{1i}}{\sqrt{2}},\sigma_{1i}\right)-\left(\frac{\mu_{2i}}{\sqrt{2}},\sigma_{2i}\right)\right\|_2}\right)^2}$$

However, the submanifold *S* is *not* totally geodesic [2], which implies that the geodesics on this submanifold *S* could be longer than the geodesics on the whole manifold. So we have:

$$D_F(\mathcal{N}(\mu_1,\Lambda_1)\|\mathcal{N}(\mu_2,\Lambda_2))\le\widetilde{D_F}(\mathcal{N}(\mu_1,\Lambda_1)\|\mathcal{N}(\mu_2,\Lambda_2))$$

which means that $\widetilde{D_F}$ is in fact an upper bound of the real geodesic distance.

Linear Dimensionality Reduction with $\mu_1\neq\mu_2$ and $r>1$: An Approximated Algorithm

Recalling that in one dimensional cases we have:

$$D_F(\mathcal{N}(\mu_1,\sigma_1^2)\|\mathcal{N}(\mu_2,\sigma_2^2))$$
$$=\sqrt{2}\sqrt{\left(\log\frac{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\|_2+\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|_2}{\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},-\sigma_2\right)\right\|_2-\left\|\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right)-\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right\|_2}\right)^2}$$

it is not difficult to realize that the above-mentioned upper bound $\widetilde{D_F}$ in high dimensional cases is a natural generalization of D_F in one dimensional cases. Since we are not able to obtain the exact closed-form solutions D_F in high dimensional cases, we could alternatively turn to maximize D_F's upper bound, $\widetilde{D_F}$.

Suppose $X\sim\mathcal{N}(0,I_n)$, $Y\sim\mathcal{N}(\mu,\Sigma)$, now our problem becomes finding a semi-orthogonal matrix $A\in\mathbb{R}^{r\times n}$ so that:

$$\widetilde{D_F}(AX\|AY)$$

is maximized.

What I used for implementation is the greedy algorithm. Heuristically, I transform this *r* dimensional problem to *r* one dimensional problems and solve them separately. First, we find a solution vector $u_1\in\mathbb{R}^{1\times n}(u_1u_1^T=1)$ which could maximize $D_F(u_1X\|u_1Y)$, just as what we have done in one dimensional cases. Then, in the subspace that is orthogonal to the solution vector u_1, we find the best solution vector $u_2\in\mathbb{R}^{1\times n}(u_2u_2^T=1)$ which could maximize $D_F(u_2X\|u_2Y)$. Afterwards, in the subspace that is orthogonal to the linear span of the solution vectors u_1,u_2, we find the best solution vector $u_3\in\mathbb{R}^{1\times n}(u_3u_3^T=1)$ which could maximize $D_F(u_3X\|u_3Y)$. We keep running the same process, until we collect r "best" solution vectors and put them together to obtain our proposed solution, *A*.

Basically we need to run the gradient ascent algorithm for *r* times to obtain the solution. Since we need to run the gradient ascent algorithm in subspaces with orthogonal constraints, still we have to make minor modifications to the algorithm and take care of certain details.

In the initialization step for the *k*-th run, we need to make sure that we start from a point which is within the subspace that is orthogonal to the linear span of the solution vectors we obtained in the previous *k*-1 runs. To do this, we first randomly generate a non-zero vector *u*, then we project *u* into the subspace by eliminating all the components that are parallel to certain solution vectors:

$$u_{k,\perp}=u-\sum_{i=1}^{k-1}<u,u_i>u_i$$

Now $u_{k,\perp}$ is orthogonal to $u_1,\dots,u_{k-1}$. Then we normalize $u_{k,\perp}$:

$$u_{k,0} = \frac{u_{k,\perp}}{\|u_{k,\perp}\|_2}$$

Hence we are able to obtain $u_{k,0}$ as the start point for the *k*-th run.

In the gradient computation step in the *k*-th run, we need to make sure that the point is updated within the subspace that is orthogonal to the linear span of the solution vectors which are obtained by the previous runs. To do this, we need to eliminate all the components of $\nabla_u^M F$ that are parallel to certain solution vectors:

$$\nabla_u^{\perp} F = \nabla_u^M F - \sum_{i=1}^{k-1} < \nabla_u^M F, u_i > u_i$$

With these modifications, we are able to implement our algorithm.

Gradient Ascent Algorithm for High Dimensional Cases

Input: μ, Σ, r

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

1. Initialize the solution vectors: $U \leftarrow \emptyset$;

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: $\nabla_{u_{k,l}}^M F \leftarrow \nabla_{u_{k,l}} F - < \nabla_{u_{k,l}} F, u_{k,l} > u_{k,l}$;

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

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

<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. Return $A = (u_1^T, \ldots, u_r^T)^T$;

Reference

[1] Shun-ichi Amari, Hiroshi Nagaoka, *Methods of Information Geometry*, Oxford University Press, 2000.

[2] Lene Theil Skovgaard, *A Riemannian Geometry of the Multivariate Normal Model*, Scandinavian Journal of Statistics, Vol. 11, No. 4 (1984), pp. 211-223.

[3] Weihuan Chen, Xingxiao Li, *An Introduction to Riemannian Geometry*, Peking University Press, 2002.

[4] Robert E. Kass, Paul W. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley Series in

Probability and Statistics, 1997.
[5] Kevin Michael Carter, *Dimensionality Reduction on Statistical Manifolds*, a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Electrical Engineering-Systems) in The University of Michigan, 2009.
[6] Sueli R. Costa, Sandra Augusta Santos, João Eloir Strapasson, *Fisher Information Matrix and Hyperbolic Geometry*, IEEE Information Theory Workshop, 2005.
[7] Kevin M. Carter, Raviv Raich, William G. Finn, Alfred O. Hero III, *Information-Geometric Dimensionality Reduction*, Signal Processing Magazine IEEE, 2011, 28(2):89-99.
[8] P. A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.