La Géométrie de l'Information: une introduction concrète et simple

M. A. G. Decastro

Résumé

La théorie moderne de la Géométrie de l'Information (ou *Information Geometry*, en anglais) est un domaine interdisciplinaire des mathématiques combinant la géométrie riemannienne, la théorie de l'information et la statistique initiée en grande partie par les travaux du mathématicien japonais Shun-ichi Amari (voir [1]). Elle suscite un certain intérêt ces dernières années du fait de ses champs d'application notamment dans le domaine de l'apprentissage machine. Tel qu'indiqué par le titre, cet article se veut une introduction pratique, simple et accessible pour un étudiant en fin de baccalauréat en mathématiques. Après un rappel de quelques outils de géométrie riemannienne, des exemples calculatoires sont présentés afin de montrer comment utiliser quelques uns des concepts majeurs de la Géométrie de l'Information dans les sections suivantes. La structure de cet article s'inspire beaucoup du livre [5].

1 Notions élémentaires de géométrie riemannienne

En Géométrie de l'Information, le cadre d'étude est la variété riemannienne. En supposant connu les notions élémentaires de topologie et sans rentrer dans les subtilités de la théorie des variétés différentielles, explicitons d'abord ce qu'on entend par variété topologique puis variété différentielle et finalement variété riemannienne en introduisant informellement et au besoin d'autres définitions.

Une variété topologique M est un espace topologique de *Hausdorff* éventuellement *à base dénombrable* et tel que pour **tout** point $p \in M$ il existe un couple (U, ϕ), appelé carte où $U \subset M, p \in U, \phi : U \longrightarrow U^{'}$ homéomorphisme et $U^{'}$ ouvert de $\mathbb{R}^n$. On appelle dimension de M le plus petit n tel que la condition précédente est vraie et *atlas* de M une famille de cartes dont les domaines U recouvrent M.

Deux cartes (U, ϕ) et (V, ψ) seront dites compatibles si $U \cap V = \emptyset$ ou bien $U \cap V \neq \emptyset$ et que *les applications de changement de cartes* $\phi \circ \psi^{-1}$ et $\psi \circ \phi^{-1}$ sont des C^k-difféomorphismes pour tout $k \geq 1$. Lorsque toutes les cartes compatibles à un atlas font déjà partie de l'atlas celui-ci est dit maximal et on le qualifie alors de structure différentiable [1].

1. On dit aussi structure différentielle

Une variété topologique devient une variété différentielle lorsqu'elle est munie d'une structure différentiable.

Un exemple standard de variété différentielle est la sphère :

$$\mathbb{S}^n = \{u = (u_1, u_2, \cdots, u_{n+1}) \in \mathbb{R}^{n+1} \mid \sqrt{u_1^2 + u_2^2 + \cdots + u_{n+1}^2} = 1\}$$

Son atlas peut être constitué uniquement de deux cartes (U, ϕ_N) et (V, ϕ_S) où $U \cap V \neq \emptyset$ et ϕ_N et ϕ_S sont des projections stéréographiques définies par :

$$\phi_N \colon \mathbb{S}^n \setminus N \longrightarrow \mathbb{R}^n \qquad \text{et} \quad \phi_S \colon \mathbb{S}^n \setminus S \longrightarrow \mathbb{R}^n$$
$$u \longmapsto \phi_N(u) = \frac{(u_1, u_2, \cdots, u_n)}{1 - u_{n+1}} \qquad u \longmapsto \phi_S(u) = \frac{(u_1, u_2, \cdots, u_n)}{1 + u_{n+1}}$$

avec $N = (1, 0, \cdots, 0)$ et $S = (0, \cdots, 0, -1)$ comme centres de projection.

Notons que les applications ϕ_N et ϕ_S sont des homéomorphismes d'inverses

$$\phi_N^{-1}(x) = \frac{\left(2x_1, 2x_2, \cdots, 2x_n, \|x\|^2 - 1\right)}{\|x\|^2 + 1} = \phi_S^{-1}(x) \quad \text{et que} \quad \phi_N \circ \phi_S^{-1}(y) = \frac{x}{\|x\|^2} = \phi_S \circ \phi_N^{-1}$$

sont bien des C^k-difféomorphismes pour tout x et y dans $\mathbb{R}^n \setminus \{0\}$.

Finalement, M est une variété riemannienne de dimension n si elle est munie d'une métrique riemannienne ou tenseur métrique, c'est-à-dire d'une forme bilinéaire [2], symétrique et définie positive g qui en tout point $p \in M$ et à tout couple (A, B) de vecteurs appartenenant à l'espace des vecteurs tangents à M au point p associe un réel $g(A, B)$. Si de plus (X, Y) est un couple de champs de vecteurs differentiables sur un ouvert contenant p alors la fonction $g(X_p, Y_p)$ est différentiable.

Les composantes de la métrique riemannienne, c'est-à-dire les entrées de la matrice associée, s'obtiennent en considérant d'abord le champs de base $(X_p^1, X_p^2, \cdots, X_p^n)$ associé à une carte puis en calculant $g_{ij} = g(X_p^i, X_p^j)$ pour $i,j = 1,2, \cdots, n$. Il y a en tout $n \times n$ fonctions composantes. Ces dernières sont utilisées pour calculer les *symboles de Christoffel* :

$$\Gamma_{ij,k} = \frac{1}{2}\left(\frac{\partial g_{jk}}{\partial \xi_i} + \frac{\partial g_{ik}}{\partial \xi_j} - \frac{\partial g_{ij}}{\partial \xi_k}\right) \quad \text{et} \quad \Gamma_{ij}^k = \sum_{l=1}^{n} \Gamma_{ij,l} g^{lk} \quad \text{avec} \quad i,j,k = 1,2, \cdots, n. \quad (1)$$

où les $\partial g_{ij}/\partial \xi_k$ sont obtenus directement par différentiation et g^{lk} correspond à la lk-ième composante de l'inverse de la métrique riemannienne.

Ces symboles (en réalité des fonctions) sont à leur tour utilisés pour calculer une *géodésique* que nous présentons au paragraphe suivant.

2. Donc ayant une représentation matricielle de dimension $n \times n$.

La géodésique est l'un des concepts les plus importants en géométrie riemannienne. Elle permet de déterminer le plus court chemin entre deux points d'une variété. Soit une courbe differentiable sur $D \subset M$, c'est-à-dire une application $c\colon I \subset \mathbb{R} \longrightarrow M$ pour laquelle on peut trouver une carte (U, ϕ) de M de champs de base associé $(X_p^1, X_p^2, \cdots, X_p^n)$ avec $U \subset D$, $p \in U$ et telle que $c \circ \phi^{-1}\colon \phi(U) \longrightarrow \mathbb{R}$ soit différentiable. Si on suppose que $g(X_p^i, X_p^j) = \delta_{ij}$ et que $\partial g_{ij(p)}/\partial X_p^i = 0$ alors il existe une application dérivée D/dt [3] telle que $\frac{D}{dt}\left(\frac{d}{dt}c(t)\right) = 0$ pour tout $t \in I$. En d'autres termes, l'accélération est nulle partout sur cette courbe. On appelle une telle courbe une géodésique. Fonctionnellement, cette dérivée seconde détermine un système d'équations différentielles du second ordre dont la (ou les) solution(s) donne(ent) exactement les fonctions composantes de la courbe du plus court chemin :

$$\frac{d^2c_k(t)}{dt^2} + \sum_{i,j=1}^{n} \Gamma_{ij}^k \frac{dc_i(t)}{dt}\frac{dc_j(t)}{dt} = 0 \quad \text{avec} \quad k = 1, 2, \cdots, n \tag{2}$$

Ici, $c_k(t)$ est la k-ième fonction coordonnée de la courbe $c(t)$. Idem pour $c_i(t)$ et $c_j(t)$.

Dans la section suivante, nous aurons l'occasion de calculer une métrique riemannienne, des *symboles de Christoffel* ainsi qu'une géodésique.

2 Métrique de Fisher et géodésiques associées

Considérons maintenant une famille de distributions normales paramétrées par deux variables, une moyenne μ et un écart-type σ :

$$S = \{p_\xi = p(x;\xi) = p_\xi(x) \mid \xi = (\mu, \sigma^2) \in \mathbb{R}\times]0, +\infty[\text{ et } x \in \mathbb{R}\}, \tag{3}$$

où p_ξ est une distribution normale de densité $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ et ξ le vecteur des paramètres μ et σ.

Puisque $\mu \times \sigma \in \mathbb{R}\times]0, +\infty[$ et qu'ils paramètrent toute la surface S alors pour tout point $p_\xi \in S$ il est raisonnable de penser qu' on peut trouver un homéomorphisme entre un ouvert contenant p_ξ et un ouvert de $\mathbb{R}\times]0, +\infty[$. De plus, la fonction de densité est infiniment différentiable par rapport aux paramètres. On peut donc supposer aisément que S est une variété différentielle où le système de coordonnées est donné par l'application $\phi(p_\xi) = \xi$.

En géométrie de l'information, une métrique riemannienne naturelle très utilisée est l'information de Fisher (voir [5, Proposition 1.7.1]) dont les composantes sont données par :

$$g_{ij}(\xi) := -E\left[\frac{\partial^2 \ln(p_\xi(x))}{\partial\xi_j\partial\xi_i}\right] = -\int_{\mathbb{R}} \frac{\partial^2 \ln(p_\xi(x))}{\partial\xi_j\partial\xi_i} p_\xi(x)dx \quad i,j = 1,2$$

3. En réalité, on appelle cette application une *connexion de Levi-Civita.*

où E est l'espérance mathématique, ln le logarithme népérien et $\partial/\partial\xi_i$ la dérivée par rapport au paramètre ξ_i.

Sachant que $\int_{-\infty}^{+\infty} p_\xi(x)dx = 1$ et en prenant $\xi_1 = \mu$ et $\xi_2 = \sigma$, on calcule aisément chacune de ses composantes :

$$
\begin{aligned}
g_{11} &= -\int_{-\infty}^{+\infty} \frac{\partial^2 \ln(p_\xi(x))}{\partial\mu^2} p_\xi(x)dx = -\int_{-\infty}^{+\infty} -\frac{1}{\sigma^2} p_\xi(x)dx = \frac{1}{\sigma^2}\int_{-\infty}^{+\infty} p_\xi(x)dx = \frac{1}{\sigma^2} \\
g_{12} = g_{21} &= -\int_{-\infty}^{+\infty} \frac{\partial^2 \ln(p_\xi(x))}{\partial\sigma\partial\mu} p_\xi(x)dx = \int_{-\infty}^{+\infty} \frac{2(x-\mu)}{\sigma^3}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&= \frac{2}{\sigma^2\sqrt{2\pi}}\int_{-\infty}^{+\infty} \frac{(x-\mu)}{\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx = -\frac{2}{\sigma^2\sqrt{2\pi}}\left[\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\Big|_{-\infty}^{+\infty}\right] = 0 \\
g_{22} &= -\int_{-\infty}^{+\infty} \frac{\partial^2 \ln(p_\xi(x))}{\partial\sigma^2} p_\xi(x)dx = -\int_{-\infty}^{+\infty} \left(\frac{1}{\sigma^3\sqrt{2\pi}} + \frac{3(x-\mu)^2}{\sigma^5\sqrt{2\pi}}\right)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&= -\frac{1}{\sigma^2} - \frac{3}{\sigma\sqrt{2\pi}}\left[\frac{-(x-\mu)}{2\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\Big|_{-\infty}^{+\infty}\right] + \frac{3}{\sigma^3\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&= -\frac{1}{\sigma^2} + \frac{3\sqrt{2\pi\sigma}}{\sigma^2\sqrt{2\pi\sigma}} = \frac{2}{\sigma^2}
\end{aligned}
$$

On peut les regrouper dans la matrice $\mathbf{g}_{ij}$ suivante dite matrice d'information de Fisher :

$$\mathbf{g}_{ij} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

Notons tout de suite que dans les calculs qui suivent k prendra les valeurs 1 et 2, en référence respectivement à μ et σ considérés à la fois comme éléments du champs de base et comme fonctions composantes.

En retenant les mêmes notations on a que :

$$\frac{\partial g_{11}}{\partial\mu} = \frac{\partial g_{12}}{\partial\mu} = \frac{\partial g_{12}}{\partial\sigma} = \frac{\partial g_{21}}{\partial\mu} = \frac{\partial g_{21}}{\partial\sigma} = \frac{\partial g_{22}}{\partial\mu} = 0, \quad \frac{\partial g_{11}}{\partial\sigma} = \frac{-2}{\sigma^3} \quad \text{et} \quad \frac{\partial g_{22}}{\partial\sigma} = \frac{-4}{\sigma^3}.$$

Et en utilisant (1) les symboles de Christoffel non-nuls sont alors :

$$\Gamma_{12,1} = \Gamma_{21,1} = -\frac{1}{\sigma^3}, \quad \Gamma_{11,2} = \frac{1}{\sigma^3}, \quad \Gamma_{22,2} = -\frac{2}{\sigma^3}$$

et

$$\Gamma^1_{12} = \Gamma_{12,1}g^{11} + \Gamma_{12,2}g^{21} = -\frac{1}{\sigma^3}\sigma^2 = -\frac{1}{\sigma}$$
$$\Gamma^1_{21} = \Gamma_{21,1}g^{11} + \Gamma_{21,2}g^{21} = \frac{1}{2}\left(-\frac{2}{\sigma^3}\right) = -\frac{1}{\sigma}$$
$$\Gamma^2_{11} = \Gamma_{11,1}g^{12} + \Gamma_{11,2}g^{22} = \frac{1}{\sigma^3}\frac{\sigma^2}{2} = \frac{1}{2\sigma}$$
$$\Gamma^2_{22} = \Gamma_{22,1}g^{12} + \Gamma_{22,2}g^{22} = -\frac{2}{\sigma^3}\frac{\sigma^2}{2} = -\frac{1}{\sigma}$$

sachant que l'inverse de la métrique $\mathbf{g}_{ij}$ est donnée par $\begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}$.

À présent, calculons la géodésique associée à l'information de Fisher. On suppose que la courbe recherchée est une fonction paramétrée par $\mu(t)$ et $\sigma(t)$. En remplaçant, dans la formule (2), les dérivées des i-ème et j-ème fonctions composantes respectivement par μ' et σ', on obtient le système d'équations différentielles du second ordre suivant :

$$\mu'' - \frac{1}{\sigma}\mu'\sigma' - \frac{1}{\sigma}\mu'\sigma' = \mu'' - \frac{2}{\sigma}\mu'\sigma' = 0 \tag{4}$$
$$\sigma'' + \frac{1}{2\sigma}\mu'^2 - \frac{1}{\sigma}\sigma'^2 = 0 \tag{5}$$

En réécrivant puis en intégrant l'équation (4) on obtient successivement :

$$\frac{\mu''}{\mu'} = 2\frac{\sigma'}{\sigma} \Longrightarrow \int \frac{\mu''}{\mu'}d\mu' = 2\int \frac{\sigma'}{\sigma}d\sigma \Longrightarrow \ln(\mu') = 2\ln(\sigma) + c \Longrightarrow \mu' = C\sigma^2 \text{ où } C = e^c \text{ et } c \text{ une constante.}$$

Deux cas sont alors possibles. Si $C = 0$, alors $\mu' = 0$ et μ est une constante K dont la dérivée dans l'équation (5) permet d'obtenir :

$$\sigma'' = \frac{1}{\sigma}\sigma'^2 \iff \frac{\sigma''}{\sigma'} = \frac{\sigma'}{\sigma} \Longrightarrow \int \frac{\sigma''}{\sigma'}d\sigma' = \int \frac{\sigma'}{\sigma}d\sigma \Longrightarrow \ln(\sigma') = \ln(\sigma) + h \Longrightarrow \ln(\sigma') - \ln(\sigma) = h$$
$$\Longrightarrow \ln\left(\frac{\sigma'}{\sigma}\right) = h \Longrightarrow \frac{\sigma'}{\sigma} = e^h \Longrightarrow \int \frac{\sigma'}{\sigma}d\sigma = \int e^h dt \Longrightarrow \ln(\sigma) = te^h + l \Longrightarrow \sigma = Le^{Ht}$$

où h et l sont des réels, $L = e^l$ et $H = e^h$.

Si, par contre $C \neq 0$, on obtient, en remplaçant μ dans l'équation (5) :

$\sigma'' + \frac{C^2}{2}\sigma^3 - \frac{1}{\sigma}\sigma'^2 = 0$. Posons $u = \sigma'$. Alors $u' = \frac{du}{dt} = \frac{du}{d\sigma}\frac{d\sigma}{dt} = \frac{du}{d\sigma}\sigma' = \frac{du}{d\sigma}u = \sigma''$.

Donc en substituant $u = \sigma'$ et $\sigma'' = u\frac{du}{d\sigma}$ dans l'équation (5) on obtient le système suivant :

$$\frac{d\sigma}{dt} = u \tag{6}$$

$$u\frac{du}{d\sigma} + \frac{C^2}{2}\sigma^3 - \frac{1}{\sigma}u^2 = 0 \tag{7}$$

L'équation (7) peut être réécrite :

$$\frac{du}{d\sigma} - \frac{1}{\sigma}u + \frac{C^2}{2}\sigma^3 u^{-1} = 0 \quad \text{avec} \quad u, \sigma \neq 0 \tag{8}$$

On obtient alors une équation de Bernoulli. Pour la résoudre, procédons comme suit : posons d'abord $z = u^2$. Alors $z' = 2uu'$ et $u' = \frac{1}{2u}z'$. En remplaçant dans l'équation (8) on obtient :

$$\frac{1}{2u}z' - \frac{1}{\sigma}u + \frac{C^2}{2}\sigma^3 u^{-1} = 0 \Longrightarrow z' - \frac{2}{\sigma}u^2 + C^2\sigma^3 = 0 \Longrightarrow z' - \frac{2}{\sigma}z = -C^2\sigma^3$$

Cette dernière équation est une équation linéaire non-homogène d'ordre 1 dont la solution générale est :

$$z = \left(k - C^2\int \sigma^3 e^{\left(-2\int \frac{1}{\sigma}d\sigma\right)}d\sigma\right)e^{2\int \frac{1}{\sigma}d\sigma} = k\sigma^2 - \frac{C^2}{2}\sigma^4 \quad \text{avec} \quad k > \frac{C^2}{2}\sigma^2$$

D'où :

$$z = u^2 \Longrightarrow u = \pm\sqrt{z} \Longrightarrow u = \pm\sqrt{k\sigma^2 - \frac{C^2}{2}\sigma^4} = \frac{d\sigma}{dt} \Longrightarrow \pm\frac{d\sigma}{\sqrt{k\sigma^2 - \frac{C^2}{2}\sigma^4}} = dt$$

$$\Longrightarrow \pm\frac{C}{\sqrt{2}}\int \frac{d\sigma}{\sigma\sqrt{\left(\sqrt{\frac{2k}{C^2}}\right)^2 - \sigma^2}} = \int dt \Longrightarrow \pm\frac{C^2}{2\sqrt{k}}\ln\left|\frac{\sqrt{\frac{2k}{C^2}} + \sqrt{\frac{2k}{C^2} - \sigma^2}}{\sigma}\right| = (t + D)$$

En posant $\frac{\sqrt{2k}}{C} = F$ et $\frac{2\sqrt{k}}{C^2} = G$ puis en prenant l'exponentielle de chaque côté on obtient [4]

$\frac{F + \sqrt{F^2 - \sigma^2}}{\sigma} = Re^{Gt}$ où $e^{GD} = R$ et, en résolvant pour σ, $\sigma = \frac{2FRe^{Gt}}{(Re^{Gt})^2 + 1}$. De plus, on sait que $\mu' = C\sigma^2$. Alors,

$$\frac{d\mu}{dt} = C\frac{4F^2(Re^{Gt})^2}{\left[(Re^{Gt})^2 + 1\right]^2} \Longrightarrow \mu = C\int \frac{4F^2(Re^{Gt})^2}{\left[(Re^{Gt})^2 + 1\right]^2}dt \text{ et donc } \mu = \frac{2CF^2}{G}\frac{-1}{(Re^{Gt})^2 + 1} + P$$

4. L'expression en valeur absolue dans l'implication précédente étant strictement supérieure à 1, on a donc laissé tomber le signe $\pm$.

avec P constante.

En résumé, selon que C soit nul ou non on obtient les géodésiques suivantes avec $\mu(t)$ et $\sigma(t)$ respectivement en première et deuxième coordonnées :

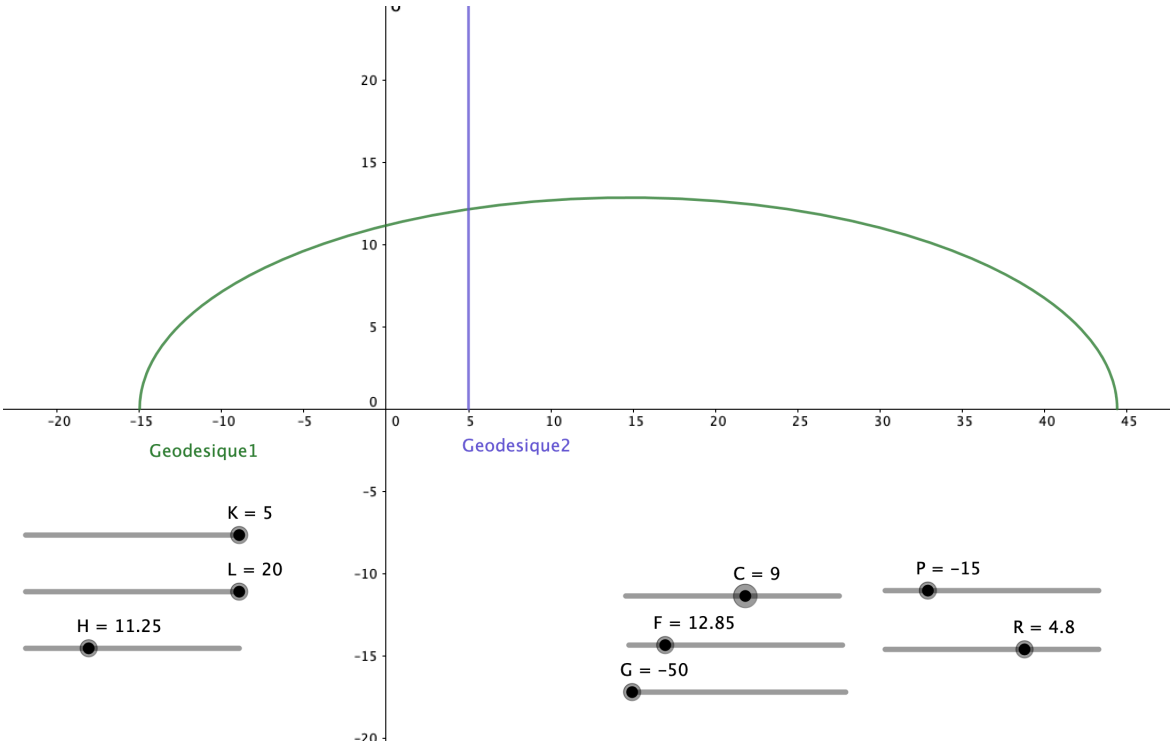
$$c(t) = \left(K, Le^{Ht}\right) \quad \text{ou} \quad c(t) = \left(\frac{\left(-2CF^2/G\right) + P(Re^{Gt})^2 + P}{(Re^{Gt})^2 + 1}, \frac{2FRe^{Gt}}{(Re^{Gt})^2 + 1} \right)$$

Dans le premier cas les géodésiques sont des demi-droites verticales (K étant une constante) alors que dans le second cas ce sont des demi-ellipses[5].

Sur Geogebra 5, nous avons écrit le code simple suivant pour visualiser les géodésiques dans des cas particuliers.

```
C = 9
F = 12.85
G = -50
P = -15
R = 4.8
K = 5
L = 20
H = 11.25
Courbe[(-2(C F²) / G + P (R e^(G t))² + P) / ((R e^(G t))² + 1),
                 (2F R e^(G t)) / ((R e^(G t))² + 1), t, -100, 100]
Courbe[K, L e^(H t), t, -100, 100]
```

5. Pour s'en rendre compte il faut voir que, pour C, F, G, P et R fixés, μ et σ satisfont l'équation d'une conique non-dégénérée $a\mu^2 + b\sigma^2 + c\mu\sigma + d\mu + e\sigma + f = 0$, avec a, b, c, d, e et f réels et a, b, c non tous nuls et dont le déterminant de la matrice $\begin{bmatrix} a & \frac{c}{2} \\ \frac{c}{2} & b \end{bmatrix}$ est strictement positif (voir [3, section 1.3].



avec μ sur l'axe des abcisses et σ sur l'axe des ordonnées.

Pour terminer la section, il est intéressant de noter, en passant, qu'avec la métrique standard sur $\mathbb{R}\times]0,+\infty[$ [6], les géodésiques seraient des demi-droites verticales et des demi-cercles.

3 Entropie de l'information et métrique de Fisher

Supposons que l'on cherche à déterminer dans (3) la distribution la plus aléatoire, c'est-à-dire celle pour laquelle on dispose le moins d'information. Clairement, cette mesure de l'incertitude devrait dépendre des paramètres. On appelle entropie une mesure du degré d'incertitude que présente une variable aléatoire et donc sa distribution de probabilité. On notera $H(p_\xi(x)) = H(\xi)$ pour signifier l'entropie associée à la distribution p_ξ. Dans le cas

6. Il est peut-être nécessaire de mentionner ici que sur $\mathbb{R}\times]0,+\infty[$, appelé demi-plan de Poincaré, il existe une métrique standard appelée métrique hyperbolique et donnée, dans le cas qui nous concerne, par $\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix}$.

où x est une variable aléatoire continue, on la definie par (voir [5, Section 3.2]) :

$$\begin{aligned} H : \mathbb{E} &\longrightarrow \mathbb{R} \\ \xi &\longmapsto -E\left[\ln(p_\xi)\right] = -\int_{-\infty}^{+\infty} p_\xi(x)\ln(p_\xi(x))dx \end{aligned} \tag{9}$$

Pour une distribution normale de paramètres μ et σ, on peut aisément calculer ce que vaut l'entropie. On utilisant (9), on a :

$$\begin{aligned} H(\xi) &= -\int_{-\infty}^{+\infty} \left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)\left(-\ln\sigma\sqrt{2\pi} - \frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \frac{\ln\sigma\sqrt{2\pi}}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx - \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} -\frac{(x-\mu)^2}{2\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \ln\sigma\sqrt{2\pi} - \frac{1}{2\sigma\sqrt{2\pi}}\left[(x-\mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx\right] \\ &= \ln\sigma\sqrt{2\pi} + \frac{1}{2} \end{aligned}$$

Observons qu'ici l'entropie ne dépend pas de μ.

À présent, essayons de répondre à la question du début de la section. Il s'agit bien évidemment d'un problème de maximisation éventuellement sous contrainte selon que des observations auront apporté des éléments d'informations supplémentaires ou non. En clair, il s'agit de déterminer les points $p \in S$ où la fonction H admet un maximum local. C'est d'abord un point critique en ce sens où la fonction H devra satisfaire :

$$\frac{\partial H(\xi)}{\partial \xi_i} = 0 \quad \text{pour} \quad i = 1,2,\cdots,n$$

Mais c'est aussi un point où le hessien $hess(H)$ de H, c'est-à-dire la matrice des dérivées partielles secondes, $\frac{\partial^2 H(\xi)}{\partial\xi_j\partial\xi_i}$, est définie négative au sens où $\langle hess(H), v\rangle < 0$[7] pour tout $v \in \mathbb{R}^n$. Les dérivées partielles secondes de H se calculent par la formule suivante (voir [5, Proposition 3.5.3]) :

$$\frac{\partial^2 H(\xi)}{\partial\xi_j\partial\xi_i} = -g_{ij}(\xi) - h_{ij}(\xi) \tag{10}$$

où $h_{ij}(\xi) = E\left[\frac{\partial\ln(p_\xi)}{\partial\xi}\frac{\partial\ln(p_\xi)}{\partial\xi_i} + \frac{\partial^2\ln(p_\xi)}{\partial\xi_i\partial\xi_j}\ln(p_\xi)\right]$.

Cette dernière équation (10) montre qu'il existe un lien entre la métrique de Fisher et l'entropie de l'information.

7. Ici, $\langle\ ,\ \rangle$ est le produit scalaire

Terminons cette section par un résultat non trivial [5, Théorème 6.5.1] : parmi toutes les distributions sur $(-\infty, +\infty)$ de moyenne μ et d'écart-type σ celle qui maximise l'entropie est la distribution normale. Si l'on se restreint au cas $\mu = 0$ et $\sigma = 1$ on voit alors que celle qui maximise l'entropie est :

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Si au lieu de travailler sur $(-\infty, +\infty)$ on travaille sur $(0, +\infty)$ on trouve une distribution exponentielle de la forme

$$p(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right)$$

et non la distribution normale !

4 Divergence de *Kullback-Leibler*

En Géométrie de l'Information, il existe des mesures de distance qui ne possèdent toutes les pas la propriété de symétrie d'une métrique. On appelle de telles mesures des *divergences* (voir [2, Section 3.2]). En particulier, une divergence sert à mesurer le "degré d'éloignement" (ou de rapprochement) de la distribution supposée (ou a priori) d'un ensemble d'observations par rapport à leur vraie loi de probabilité. Une divergence très utilisée est la divergence de *Kullback-Leibler* aussi appelée entropie relative. Dans la suite, on se restreindra aux cas des variables aléatoires continues.

Soit $p_{\xi_p}(x)$ et $q_{\xi_q}(x)$ deux distributions continues de vecteurs de paramètres respectivement $\xi_p = (\mu_p, \sigma_p)$ et $\xi_q = (\mu_q, \sigma_q)$. On définit (voir [5]) la divergence de *Kullback-Leibler*, D_{KL}, entre les distributions $p_\xi(x)$ et $q_\xi(x)$ par :

$$D_{KL}(p||q) = E_{p_{\xi_p}}\left[\ln \frac{p_{\xi_p}(x)}{q_{\xi_q}(x)}\right] = \int_{-\infty}^{+\infty} p_{\xi_p}(x) \ln \frac{p_{\xi_p}(x)}{q_{\xi_q}(x)} dx$$

On peut illustrer le calcul de la divergence de *Kullback-Leibler* pour deux distributions $p_{\xi_p}(x) = \frac{1}{\sigma_p\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)$ et $q_{\xi_q}(x) = \frac{1}{\sigma_q\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)$ sur la variété S défine par (3) où $q_{\xi_q}(x)$ comme la distribution a priori et $p_{\xi_p}(x)$ comme la vraie distribution. On a :

$$
\begin{aligned}
D_{KL}(p||q) &= \int_{-\infty}^{+\infty} p_{\xi_p}(x) \ln \left(\frac{\frac{1}{\sigma_p\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)}{\frac{1}{\sigma_q\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)} \right) dx \\
&= \int_{-\infty}^{+\infty} p_{\xi_p}(x) \left(-\ln \sigma_p\sqrt{2\pi} - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \ln \sigma_q\sqrt{2\pi} + \frac{(x-\mu_q)^2}{2\sigma_q^2} \right) dx \\
&= \ln \frac{\sigma_q}{\sigma_p} - \int_{-\infty}^{+\infty} \frac{(x-\mu_p)^2}{2\sigma_p^2} p_{\xi_p}(x)dx + \int_{+\infty}^{-\infty} \frac{(x-\mu_q)^2}{2\sigma_q^2} p_{\xi_p}(x)dx \\
&= \ln \frac{\sigma_q}{\sigma_p} - \frac{1}{2} + \int_{+\infty}^{-\infty} \frac{(x-\mu_q)^2}{2\sigma_q^2} p_{\xi_p}(x)dx
\end{aligned}
$$

La troisième intégrale s'obtient en posant $\Delta\mu = \mu_p - \mu_q$ et en faisant les transformations nécessaires. D'où le résultat :

$$
D_{KL}(p||q) = \ln \frac{\sigma_q}{\sigma_p} - \frac{1}{2} + \frac{\sigma_p^2 + (\Delta\mu)^2}{2\sigma_q^2}
$$

Comme il fallait s'y attendre, lorsque les paramètres ξ_q tendent vers ξ_p, $D_{KL}(p||q)$ tend vers zéro et la dérivée

$$
\frac{\partial}{\partial \xi_i} D_{KL}(p||q)
$$

y vaut également zéro.

Terminons par une relation qui montre le lien entre la divergence de *Kullback-Leibler* et la métrique de Fisher. C'est celle-ci :

$$
\left. \frac{\partial^2}{\partial \xi_i \partial \xi_j} \right|_{\xi_q = \xi_p} D_{KL}(p||q) = g_{ij}(\xi_p) \quad \text{pour} \quad i,j = 1,2,\cdots,n
$$

où $g_{ij}(\xi_p)$ est la ij-ème composante de la métrique $\mathbf{g}_{ij}$ au point $p_{\xi_p} = (\mu_p, \sigma_p)$ et l'égalité $\xi_q = \xi_p$ signifie que la dérivée seconde est évaluée en (μ_p, σ_p). Dans le cas de notre variété S, on a voit bien que :

$$
\frac{\partial^2 D_{KL}(p||q)}{\partial \mu_p^2} = \frac{1}{\sigma_q^2} = g_{11},\ \frac{\partial^2 D_{KL}(p||q)}{\partial \sigma_p \partial \mu_p} = \frac{\partial^2 D_{KL}(p||q)}{\partial \mu_p \partial \sigma_p} = 0 \text{ et } \frac{\partial^2 D_{KL}(p||q)}{\partial \sigma_p^2} = \frac{2}{\sigma^2} = g_{22}
$$

Références et bibliographie

[1] S. Amari. *Differential-geometrical Methods in Statistics*. Lecture notes in Statistics. Springer-Verlag, 1985.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*. Translations of mathematical monographs. American Mathematical Society, 2000.

[3] D.A. Brannan, D.A. Brannan, M.F. Esplen, and J.J. Gray. *Geometry*. Cambridge University Press, 1999.

[4] P. Buser. Notes "inofficielles" du cours de géométrie riemannienne. https://moodlearchive.epfl.ch/2018-2019/mod/resource/view.php?id=881163, Automne 2010. École Polytechnique Fédérale de Lausanne, consulté le 17 septembre 2021.

[5] O. Calin and C. Udrişte. *Geometric Modeling in Probability and Statistics*. Mathematics and Statistics. Springer International Publishing, 2014.

[6] Sueli I.R. Costa, Sandra A. Santos, and João E. Strapasson. Fisher information distance : A geometrical reading. *Discrete Applied Mathematics*, 197 :59–69, 2015. Distance Geometry and Applications.

[7] J. Lafontaine. *An Introduction to Differential Manifolds*. Springer International Publishing, 2015.

[8] S. Lipschutz. *Schaum's Outline of General Topology*. Schaum's Outline Series in Mathematics. McGraw-Hill Companies,Incorporated, 1965.

[9] Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10) :1100, Sep 2020.

[10] A.N. Pressley. *Elementary Differential Geometry*. Springer Undergraduate Mathematics Series. Springer London, 2010.

[11] L.W. Tu. *An Introduction to Manifolds*. Universitext. Springer New York, 2010.