Fouille de Données et Media Sociaux

Cours 5 – Géométrie riemannienne pour l'apprentissage

Olivier Schwander <olivier.schwander@lip6.fr>

Master DAC Data Science
UPMC - LIP6

Lundi 15 octobre 2018

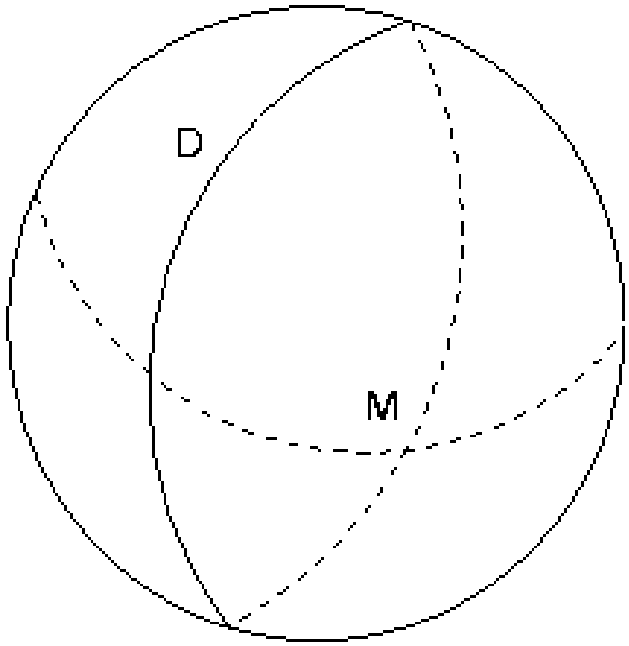
Géométrie euclidienne

1. Un segment de droite peut être tracé en joignant deux points quelconques distincts ;
2. Un segment de droite peut être prolongé indéfiniment en une ligne droite ;
3. Étant donné un segment de droite quelconque, un cercle peut être tracé en prenant ce segment comme rayon et l'une de ses extrémités comme centre ;
4. Tous les angles droits sont congruents ;
5. Si deux lignes sont sécantes avec une troisième de telle façon que la somme des angles intérieurs d'un côté est strictement inférieure à deux angles droits, alors ces deux lignes sont forcément sécantes de ce côté.

Somme des angles d'un triangle: 180°

Géométrie elliptique

Par un point extérieur à une droite, aucune droite parallèle ne passe.

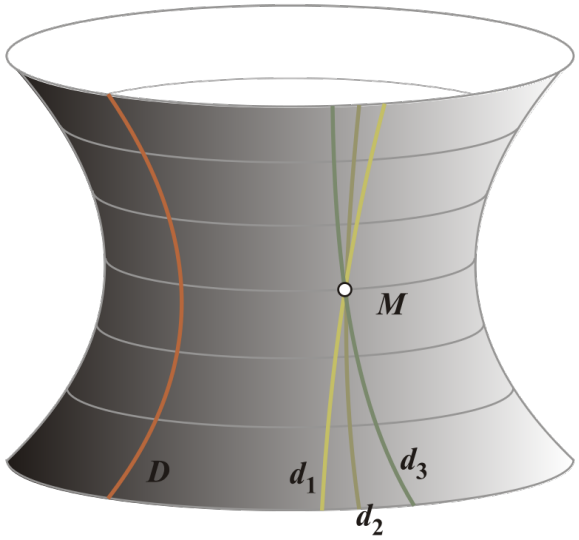


Somme des angles d'un triangle: $> 180°$

Source: https://commons.wikimedia.org/wiki/File:Para_riemann.png

Géométrie hyperbolique

Par une point extérieur à une droite, il y a une infinité de droites parallèles.

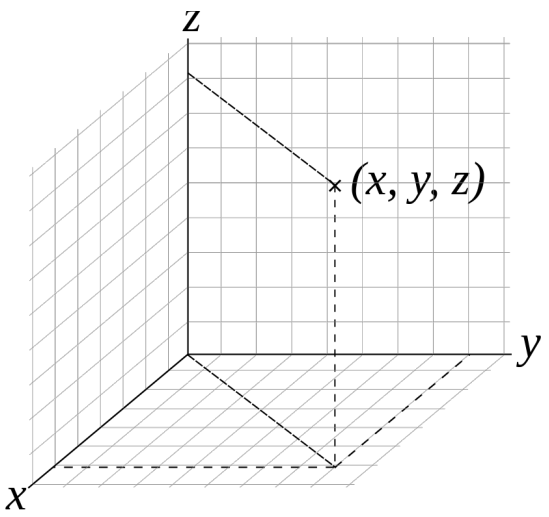


Somme des angles d'un triangle: $< 180°$

Source: https://commons.wikimedia.org/wiki/File:LineOnHyper.svg

Systèmes de coordonnées

Géométrie analytique



Pas unique ! (autres repères cartésiens, coordonnées polaires, etc)

Source: https://commons.wikimedia.org/wiki/File:3D_Cartesian.svg; Le trésor de Rackham le Rouge, Hergé

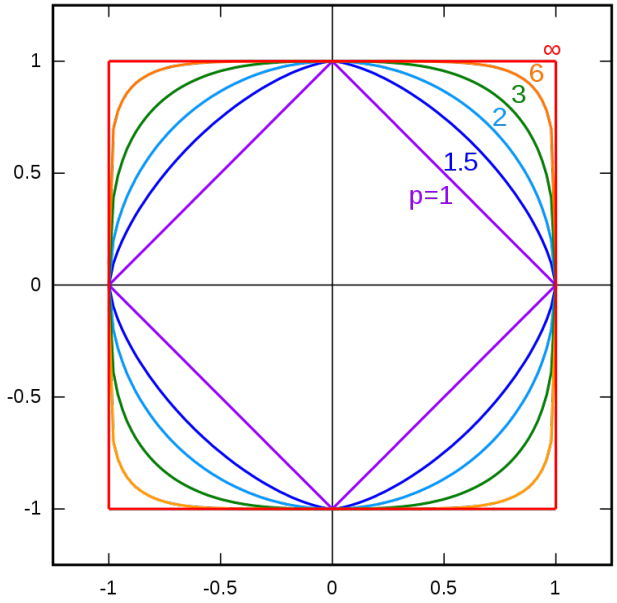
Distance

Distance euclidienne

$$d_2(x,y) = ||x-y||_2 = \left(\sum_{i=1}^{d} |x_i - y_i|^2\right)^{\frac{1}{2}}$$

Distance**S**

$$d_p(x,y) = ||x-y||_p = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$



Et plein d'autres (pondérations ou Mahalanobis par exemple)

Source: https://commons.wikimedia.org/wiki/File:Vector-p-Norms_qtl1.svg

Géométrie, distance et coordonnées

À quoi correspond géométriquement la distance euclidienne ?

Est-ce que ça dépend du système de coordonnées ? Et pour les autres distances ?

Et avec des coordonnées polaires ?

Barycentre

Centre de gravité: le point O tel que $m_1\overrightarrow{OA} + m_2\overrightarrow{OB} = 0$

$$O = \frac{m_1}{m_1 + m_2}\overrightarrow{AB}$$

Est-ce que ça dépend du système de coordonnées ? D'une notion de distance ? De la géométrie ?

Source: https://commons.wikimedia.org/wiki/File:Barycentre1.png

BarycentreS

1. Par axiomatisation (Kolmogorov 1930, Nagumo 1930, Aczél 1948)
   - Réfléxivié: $B(x, x) = x$
   - Symmétrie: $B(x, y) = B(y, x)$
   - Continuité et stricte monotonicité
   - Associativité: $B(B(x, y), z) = B(x, B(y, z))$

   Entre autre: moyenne arithmétique, géométrique, harmonique

2. Par optimisation

$$B(x_1, \dots, x_N) = \arg\min_{x^*} \sum_i d^2(x^*, x_i)$$

Entre autre: moyenne arithmétique, géométrique, harmonique

Géométrie de l'information

Étude géométrique des distributions de probabilité

- Distance entre distributions, ou autres dissimilarités
- Barycentre
- Plus court chemin

Objets géométriques: points sur la variété statistique

Système de coordonnées: paramètres des distributions

Géométrie de l'information algorithmique: comment calculer (efficacement) tout ça

Distance de Fisher-Rao-Hotelling

Matrice d'information de Fisher

$$I_{ij}(\theta) = E\left[\frac{\partial \log p_\theta}{\partial \theta_i}\frac{\partial \log p_\theta}{\partial \theta_i}\right] = E\left[\frac{1}{p_\theta}\frac{\partial p_\theta}{\partial \theta_i}\frac{1}{p_\theta}\frac{\partial p_\theta}{\partial \theta_i}\right]$$

Métrique

$$\mathrm{d}s = \mathrm{d}s^2(\theta) = \sum_{i=1}^{D}\sum_{j=1}^{D} I_{ij}(\theta) d\theta_i d\theta_j = (\nabla\theta)^T I(\theta)\nabla\theta$$

Distance

$$\mathrm{FR}(\theta_1, \theta_2) = \min_{\theta(t) \text{ st } \theta(0)=\theta_1, \theta(1)=\theta_2} \int_0^1 \sqrt{ds^2}dt$$

Interprétation et propriété

- Métrique: notion de distance locale (une distance de Mahalanobis)
- Distance: plus court chemin sur la variété statistique

Propriété d'invariance

- Invariance par reparamétrisation
- Indépendant du système de coordonnées
- Pas de biais lié à un choix arbitraire
- Unique notion de distance avec cette propriété (Censov 1982, Hông Vân Lê 2016)

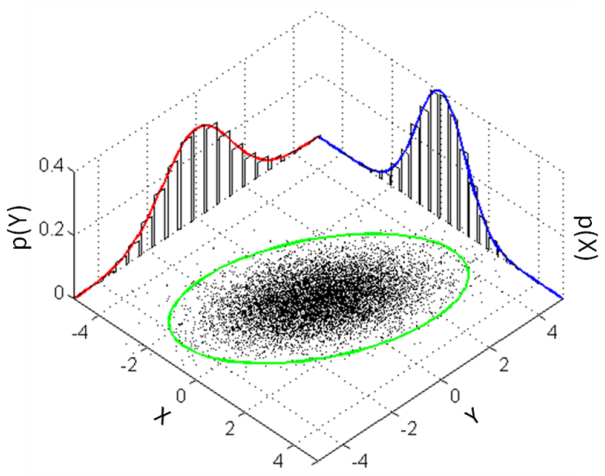
Rarement en forme close:

- loi multinomiale: géométrie sphérique
- loi normale: géométrie hyperboliquqe

Gaussiennes multivariées

Pour une matrice Σ symétrique positive définie positive:

$$p(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right)$$



Pour $\mu = 0$, distance de Fisher-Rao:

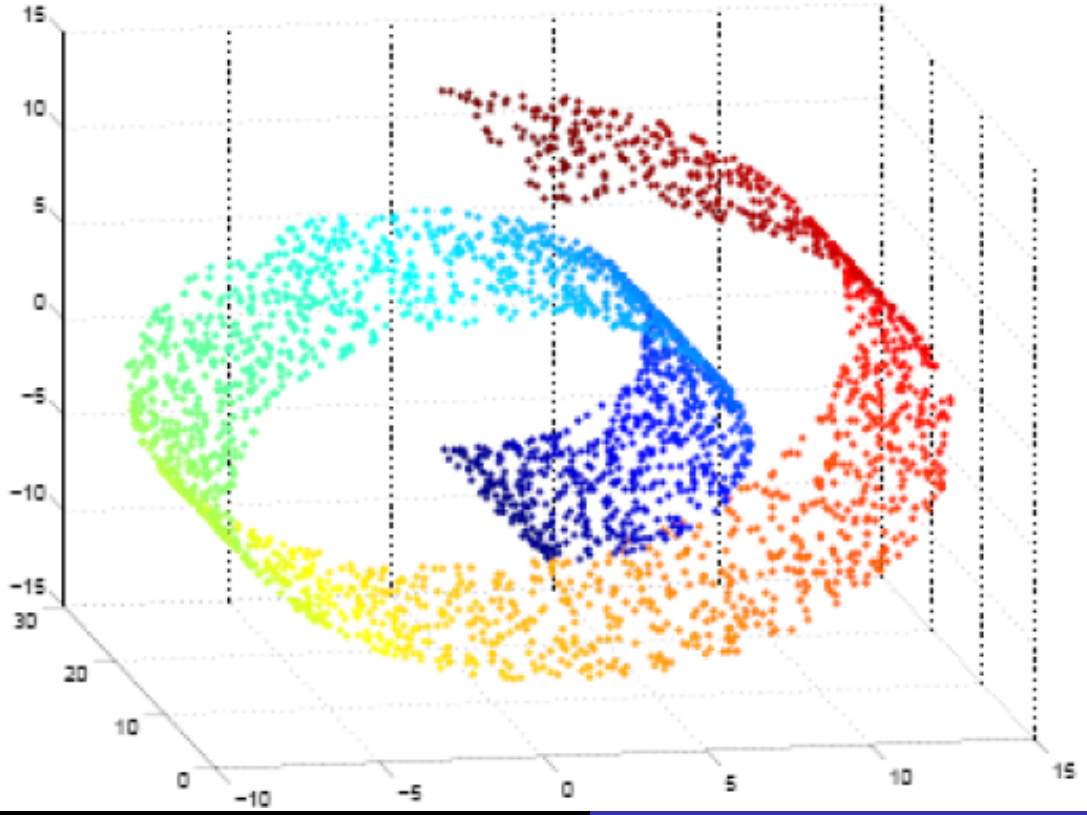
$$\delta(P_1, P_2) = \frac{1}{2}||log(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}})||_F$$

Source: https://en.wikipedia.org/wiki/File:MultivariateNormal.png

Variété

- Courbe: variété de dimension 1
- Surface: variété de dimension 2

Définition Une variété de dimension n est un espace topologique *localement euclidien*



Atlas

Ensemble de cartes locales

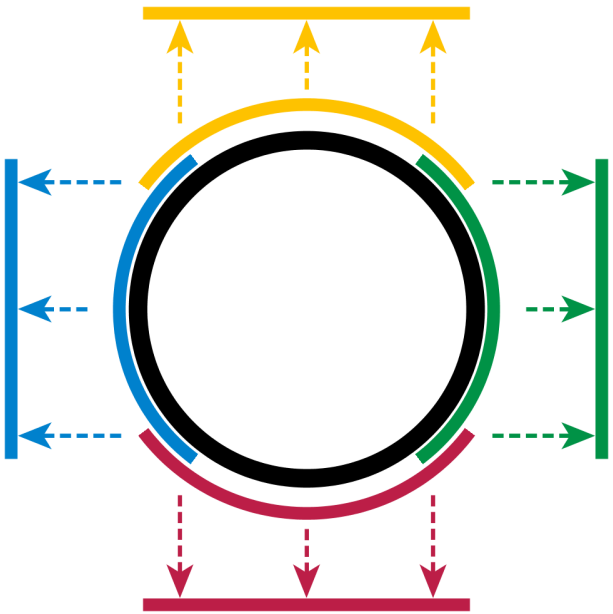
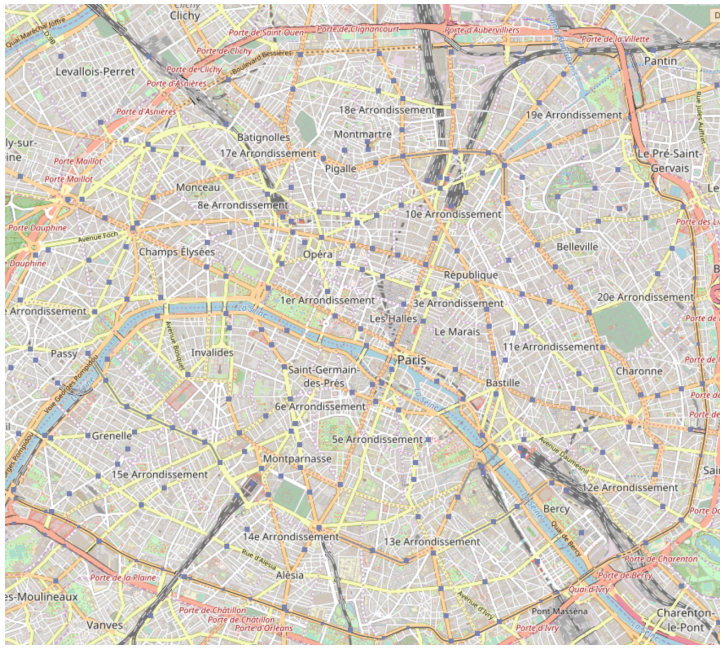


Figure Wikipédia

Système de coordonnées locales

Atlas

The Blue Marble / NASA

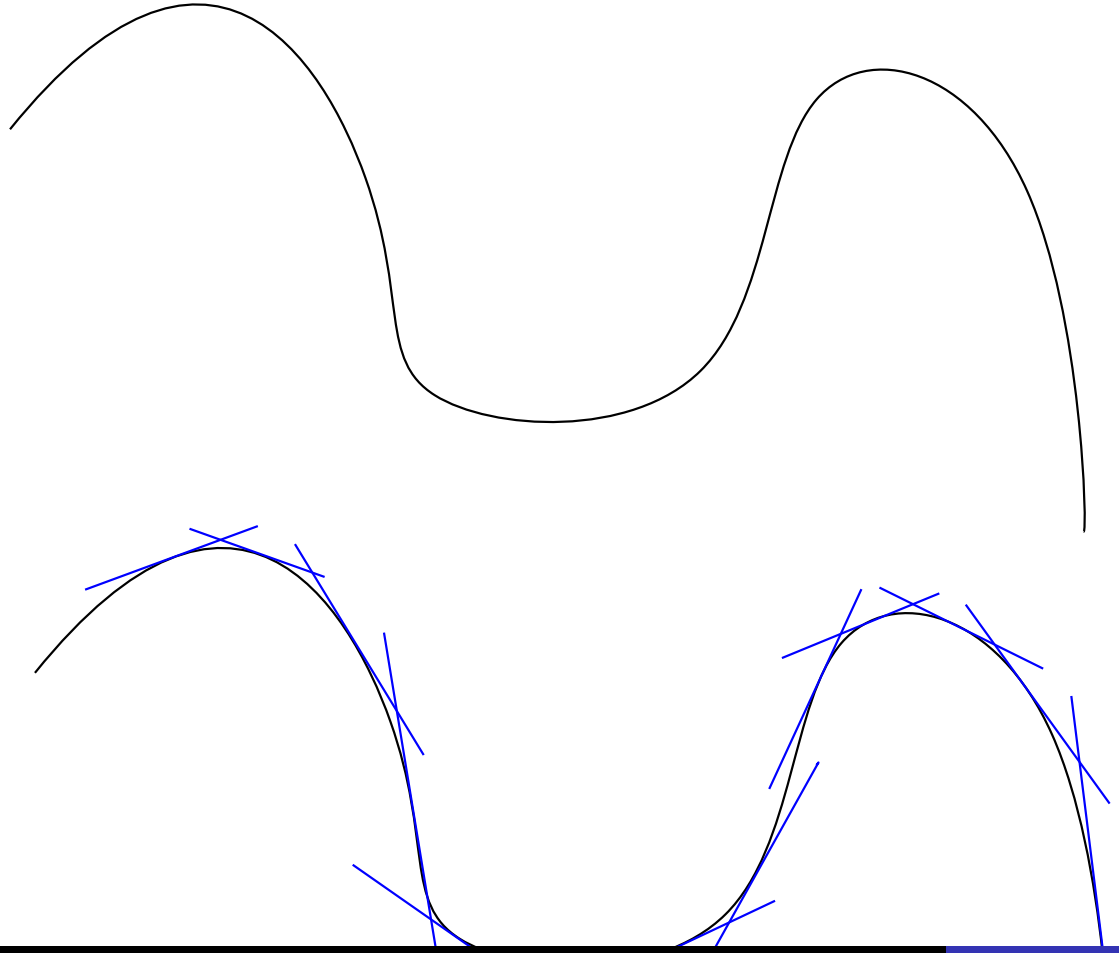


Openstreetmap

Variété riemannienne

Variété différentielle, avec une structure métrique

Atlas: espaces tangents, chacun avec un produit scalaire local



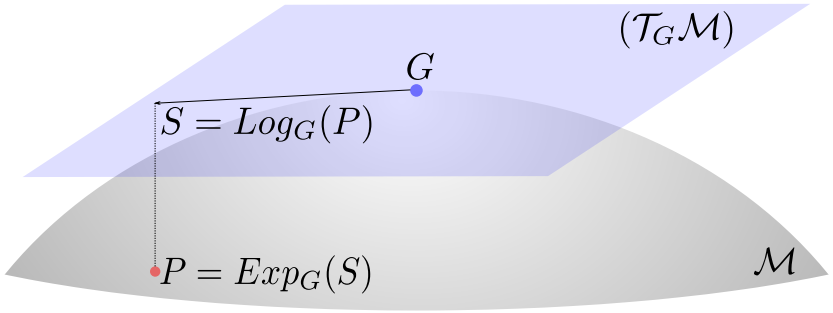
Espaces tangents

Système de coordonnées locales autour de G

- G origine des coordonnées

Projections

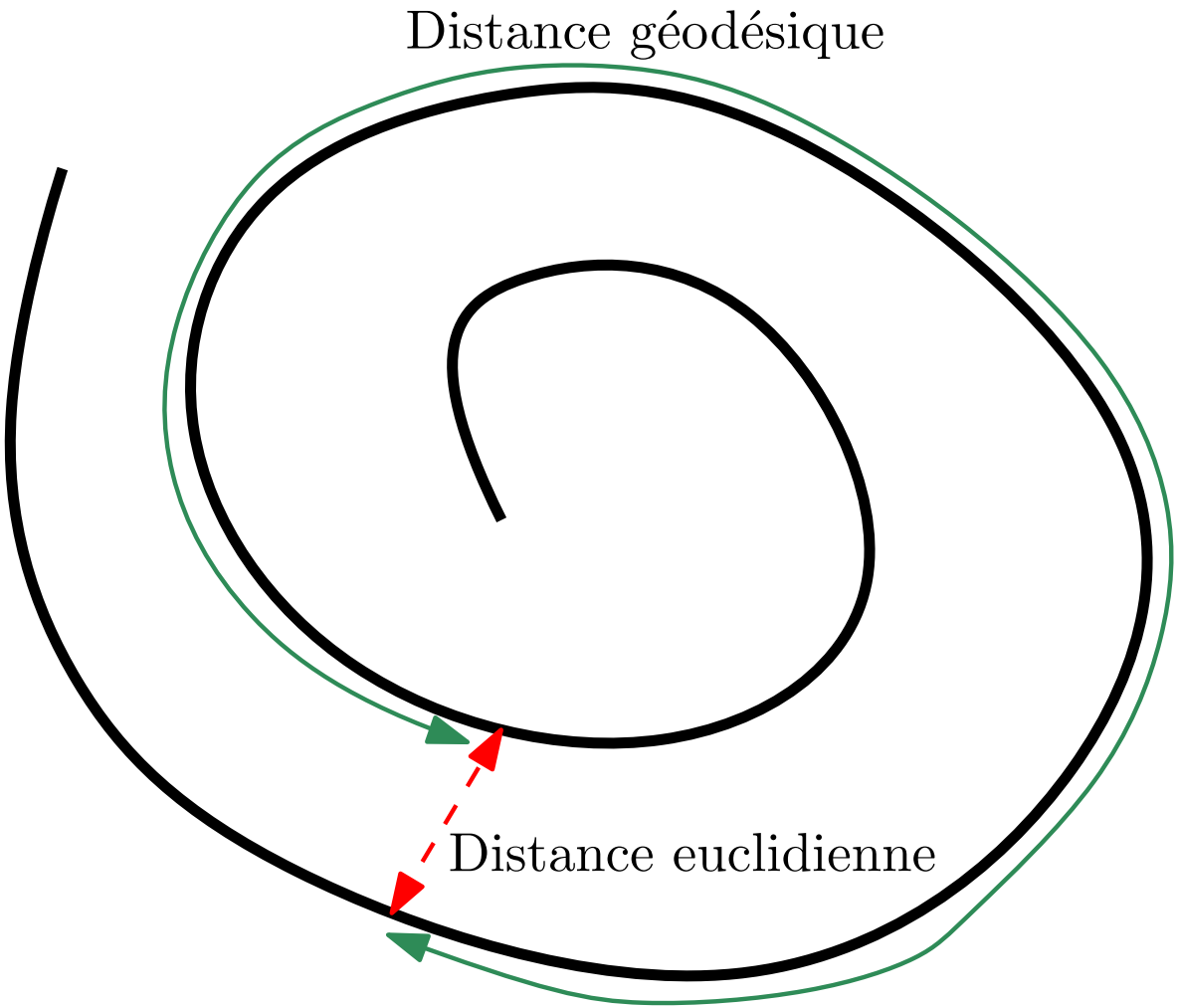
- logarithmique: variété vers hyperplan
- exponentielle: hyperplan vers variété



Source: Daniel Brooks

Distance géodésique

- Plus cours chemin sur la variété



Distance riemannienne

Sur chaque espace tangent:

- produit scalaire local
- métrique locale $\mathbf{F}_P$

$$\delta(P_1, P_2) = \inf_{\xi \mid (\xi(0)=P_1, \xi(1)=P_2)} \int_0^1 ds(t)dt$$

$$ds(t)^2 = \dot{\xi}(t)^T \mathbf{F}_{\xi(t)} \dot{\xi}(t)$$

Barycentre riemannien

Par optimisation

$$B(P_1, \dots, P_N) = \arg\min_{P^*} \sum_i \delta^2(P^*, P_i)$$

Pas de forme close en général

Cas des matrices de covariances

ou des gaussiennes multivariées à moyenne nulle

Distance:

$$\delta(P_1, P_2) = \frac{1}{2}||log(P_1^{-\frac{1}{2}} P_2 P_1^{-\frac{1}{2}})||_F$$

La variété statistique est une variété riemannienne: les deux points de vue coïncident.

Propriétés de la distance

- Invariance par inversion

$$\delta(P_1, P_2) = \delta(P_1^{-1}, P_2^{-1})$$

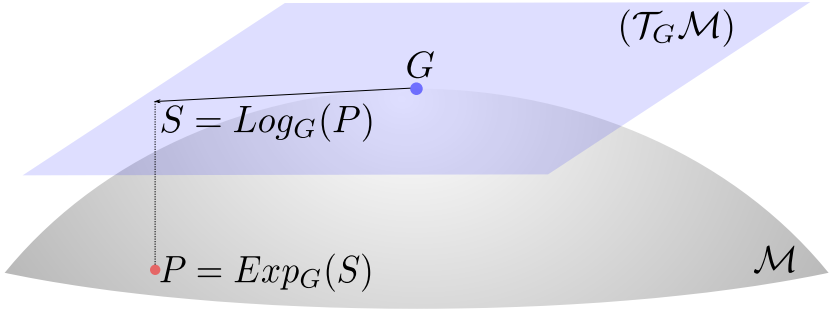
- Invariance par congruence

$$\delta(P_1, P_2) = \delta(P^t P_1 P, P^t P_2 P)$$

Conséquences de l'invariance par reparamétrisation !

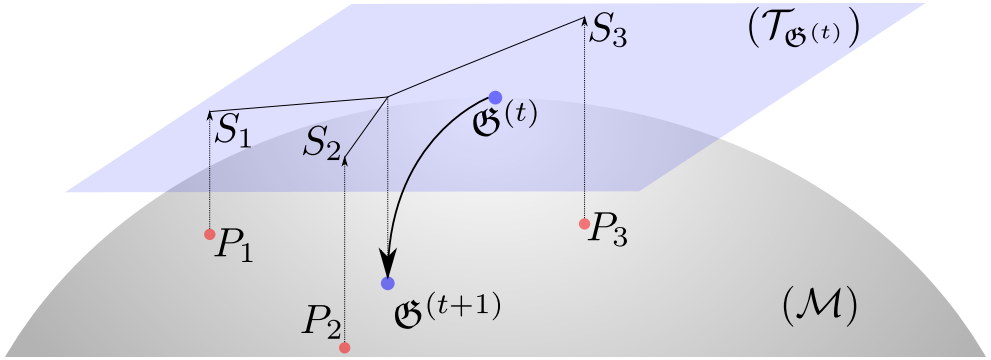
Projections pour les matrices de covariances

$$\forall S \in \mathcal{T}_G, Exp_G(S) = G^{\frac{1}{2}} exp(G^{-\frac{1}{2}} S G^{-\frac{1}{2}}) G^{\frac{1}{2}} \in \mathcal{S}_*^+$$
$$\forall P \in \mathcal{S}_*^+, Log_G(P) = G^{\frac{1}{2}} log(G^{-\frac{1}{2}} P G^{-\frac{1}{2}}) G^{\frac{1}{2}} \in \mathcal{T}_G$$



Barycentre pour les matrices de covariances

Barycentre itératif: flot de Karcher (1977)



Source: Daniel Brooks

Matrice de covariance

Covariance, pour un vecteur $X = (x_1, \dots, x_d)$

$$\Sigma = \mathrm{cov}(X, X) = \mathrm{E}\big[(X - \mathrm{E}[X])(X - \mathrm{E}[X])\big]$$

- Mesure du lien statistique entre les variables
- Variables non-corrélées: matrice diagonale

Covariance comme représentation

Détection d'objets:

- Tuzel, Porikli, and Meer 2006
- Covariance d'une région, sur une feature map
- Images intégrales, descripteurs à la Viola-Jones
- Distance Fisher Rao

Imagerie médicale (Diffusion Tensor Imaging):

- Pennec, Fillard, and Ayache 2006
- Dispositif physique qui renvoie un tenseur de matrices SPD 3x3
- Moyennes de représentations du cerveau
- Calculs efficaces des distances, des moyennes

Covariance empirique

Sur une échantillon $X_1, \dots, X_N$:

$$P = \frac{1}{N-1} \sum_i X_i X_i^t$$

Estimateur non-biasé sur $N \gg d$

Propriété importante: additivité

Estimateurs robustes

Version régularisée:

$$P^+ = P + \lambda tr(P) I$$

Séries temporelles multivariées

Données d'entrée:

$$X_i = \begin{pmatrix} x_1^{(1)} & x_1^{(1)} & \dots & x_{T_i}^{(1)} \\ \dots & & & \\ x_1^{(d)} & x_1^{(d)} & \dots & x_{T_i}^{(d)} \end{pmatrix}$$

- Taille variable
- Description globale: une seule matrice de covariance

Électro-encéphalogrames

Séries temporelles multivariées:

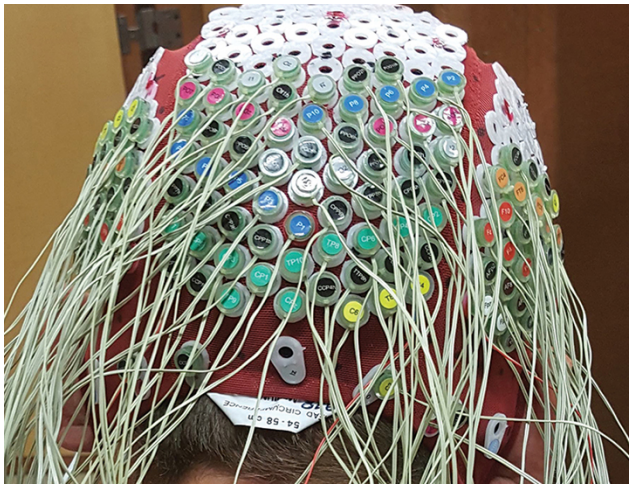
- Barachant et al. 2013
- Covariance sur les différentes variables
- Distance Fisher-Rao et plus proche voisin

État de l'art pour l'interface cerveau-machine

Interface cerveau-machine

- Chaque variable reçoit un mélange des sources d'entrées
- Fisher-Rao: invariance par ce type de transformation !

Besoin d'une robustesse au faible nombre d'exemples d'apprentissage: en fait, on préfère apprendre le moins possible



Source: https://www.medgadget.com/2017/12/new-eeg-system-achieves-higher-signal-resolution-ever.html

Vidéos

Série temporelle d'images:

- Variables: tous les pixels
- Covariance globale
- Puis réduction de dimension

Reconnaissance d'actions, d'émotions:

- Huang and Van Gool 2017
- Réseau de neurones adapté pour les matrices SPD

Plus proche voisin

Plus proche voisin classique: mais avec la distance de Fisher-Rao

- Pas d'apprentissage

Pratique pour l'interface cerveau-machine

Plus proche barycentre

Apprentissage:

- Calcul du barycentre pour chaque classe

Bon compromis pour l'interface cerveau-machine

SVM

Recherche d'un séparateur linéaire qui maximise la marge

- Besoin d'un espace euclidien
- Donc dans un espace tangent

Lequel choisir ?

- Identité
- Barycentre

Clustering

k-means

- Distance Fisher-Rao
- Barycentre Fisher-Rao

Convergence:

- Barycentre défini comme une minimisation du compromis

Retour sur les réseaux de neurones classiques

Étage de représentation

- Transformation des données (couches denses, convolutions)
- Non-linéarités

Étage de classification

- Classifieur non-linéaire (MLP)
- Sorties adaptées à la tâche (classification, régression, segmentation)

Remarque sur les convolutions: comme du dense, mais avec plus de structure

SPDNet (Huang and Van Gool 2017): adaptation d'un réseau de neurones pour les matrices SPD

Couche de transformation: BiMap

Matrice d'entrée de taille $n \times n$

- Transformation linéaire de la matrice SPD

$$P^{(l)} = W^{(l)^T} X^{(l-1)} W^{(l)}$$

avec $W^{(l)}$ semi-orthogonale (variété Stieffel), de taille $n \times m$

En sortie: matrice SPD de taille $m \times m$

- Correspond à un changement de base et une troncature

Couche de non-linéarité: ReEig

- Activation sur les valeurs propres

$$X^{(l)} = U^{(l)} \max(\Sigma^{(l)}, \epsilon I_n) U^{(l)^T}$$

avec $P^{(l)} = U^{(l)} \Sigma^{(l)} U^{(l)^T}$

- Sorte de ReLU
- Sorte de régularisation: évite les valeurs propres trop faibles

Étage de classification: LogEig

- Classification finale: MLP

Besoin d'un espace euclidien

Projection logarithmique sur un espace tangent

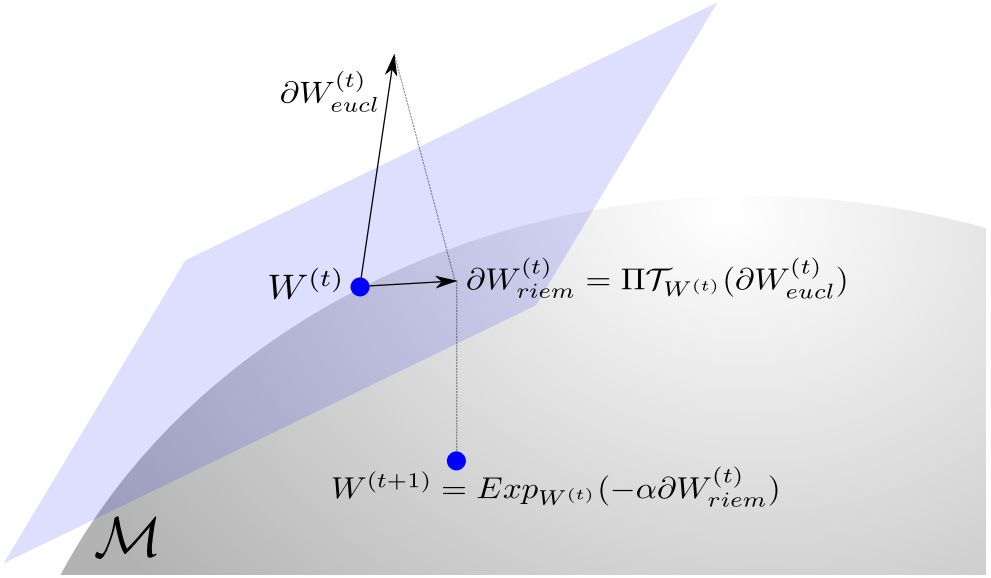
- Lequel choisir ?
- Matrice identité

Descente de gradient avec contraintes

Descente de gradient classique: impossible

- Matrices structurées
- Respect des contraintes

$$\Pi\mathcal{T}_W(X) = X - WW^TX$$
$$Exp_W(X) = Orth(W + X)$$



Source: Daniel Brooks

Applications du SPDNet

Papier original:

- Vidéo: reconnaissance d'action, d'émotions

MLIA:

- Classification de signaux radar (drones)
- Classification de signaux audio

CNN + SPDNet

Matrice de covariance:

- Données brutes
- Représentation
- Feature map d'un CNN pré-appris

End-to-end:

- Brancher le SPDNet après un CNN
- Backprop classique

Meilleur LogEig

Projection par rapport à l'identité

- Ne dépend pas des données

Autres choix:

- Point de référence: barycentre
- Point de référence appris (encore descente de gradient projeté, sur la variété des matrices SPD)

Conclusion

Méthodes originales:

- Au moins en machine learning
- Bonnes fondations mathématiques et statistiques
- Relativement bien connu de la communauté signal

Intérêt:

- Mieux exploiter le structure des données
- Contextes avec peu de données
- Audio (relativement peu de données, CNN pas suffisants)