Structure géométrique de modèles statistiques

Germain Van Bever

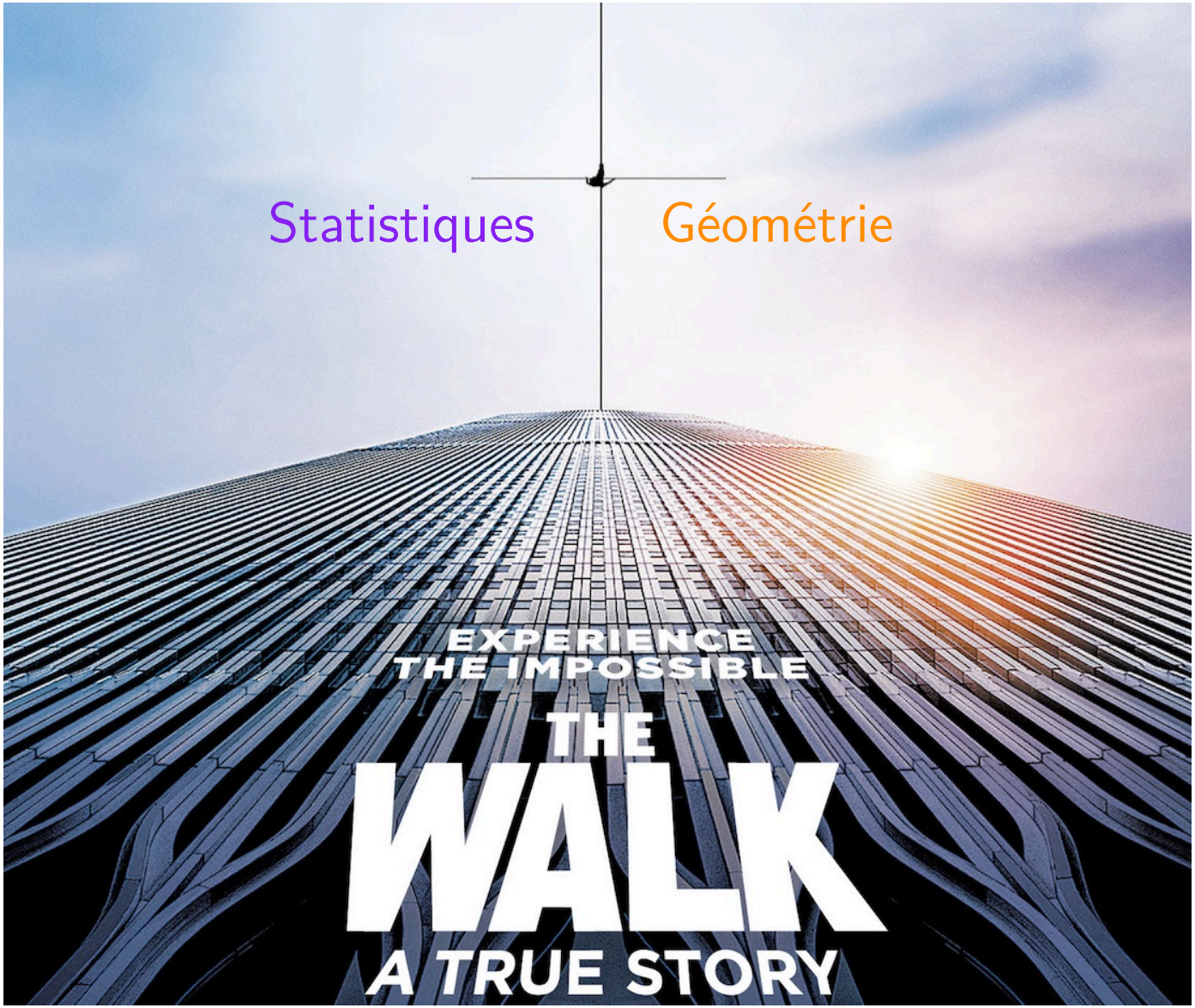
The Open University (il y a 14h01)
Université libre de Bruxelles / FNRS (depuis 14h00)

BSSM, 1er août 2016

Introduction

Introduction

Introduction



Motivation

Dans cet exposé, nous allons

- considérer un ensemble de distributions de probabilité comme une variété,
- analyser la structure géométrique de cette variété et explorer son lien avec l'estimation statistique, via
- les concepts de métrique, connexions affines, géodésiques, courbure, etc.

Table des matières

Table des matières

Table des matières

Modèle statistique

- Un modèle statistique est un ensemble de lois de probabilités sur $\mathcal{X}$:

$$\mathcal{P} \subset \mathcal{P}_{\mathcal{X}} := \left\{ p : \mathcal{X} \to \mathbb{R} \middle| \; p(x) > 0 \; (\forall x \in \mathcal{X}), \int p(x)dx = 1 \right\}.$$

- Un modèle $\mathcal{P}$ est dit *paramétrique* si

$$\mathcal{P} = \{ p = p(\cdot; \xi) | \xi \in \Xi \subset \mathbb{R}^n \}.$$

- On supposera le modèle considéré *régulier*.

Exemple: les lois normales

$\mathcal{X} = \mathbb{R}$, $n = 2$, $\xi = (\mu, \sigma)$, $\Xi = \{(\mu, \sigma) | -\infty < \mu < \infty, 0 < \sigma < \infty\}$

$$p(x; \xi) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right).$$

Variété

Définition (vague)

Une variété est un "ensemble S muni d'un système de coordonnées", i.e. une bijection (d'un sous-ensemble) de S dans (un sous-ensemble ouvert de) $\mathbb{R}^n$.

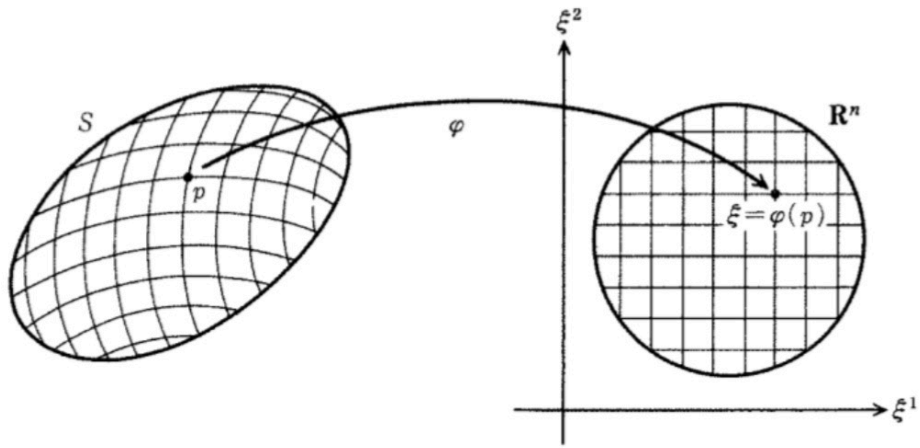


Figure: Un système de coordonnées ξ sur S.

Nous supposerons, sans perte de généralité, avoir un système de coordonnées global.

Par abus de langage, on dira souvent "l'élément ξ de S".

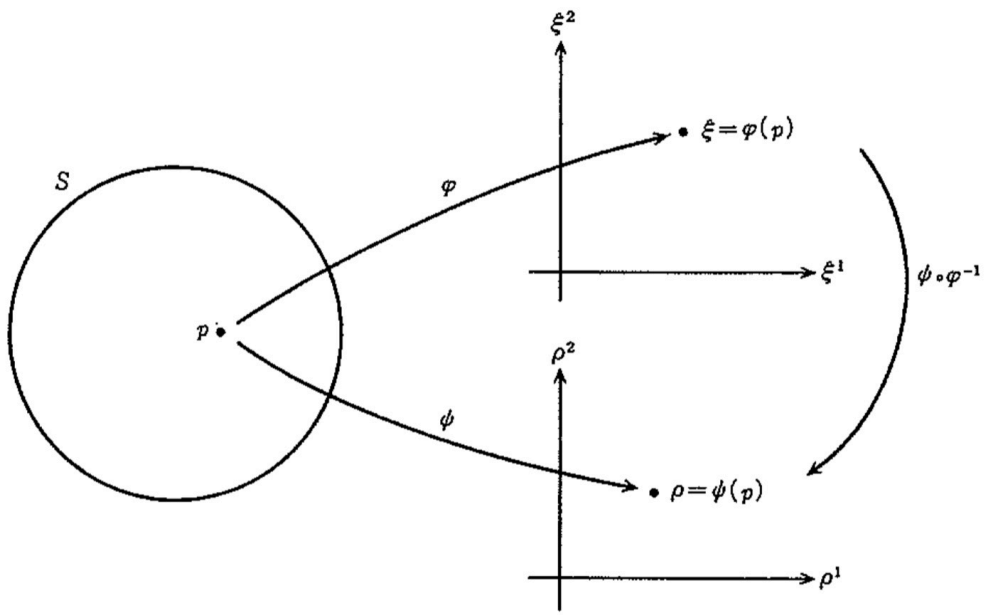
Variété différentielle

Définition

Une variété différentielle est un ensemble S pour lequel il existe un ensemble de systèmes de coordonnées $\mathcal{A}$ satisfaisant les conditions

1. *Chaque élément $\phi \in \mathcal{A}$ est une bijection de S dans $\mathbb{R}^n$.*
2. *Pour tout $\phi \in \mathcal{A}$ et pour toute bijection $\psi : S \to \mathbb{R}^n$, on a*

$$\psi \in \mathcal{A} \Leftrightarrow \psi \circ \phi^{-1} \text{ est un difféomorphisme } C^\infty.$$



Exemples

- En dimension 1: Une droite dans $\mathbb{R}^n$, un cercle, tout sous-ensemble ouvert de ceux-ci.
- En dimension 2: La surface d'une sphère, d'un tore, d'un hémisphère.
- En dimension 3: Espaces de couleurs

Famille exponentielle

Définition

Un modèle paramétrique est une famille exponentielle si

$$p = p(x;\theta) = exp\left(C(x) + \sum_{i=1}^{n} \theta_i F_i(x) - \psi(\theta)\right),$$

pour θ_i *sont les* paramètres naturels, ψ *la fonction potentielle, satisfaisant*

$$\psi(\theta) = log \int exp\big(C(x) + \sum_{i=1}^{n} \theta_i F_i(x)\big) dx.$$

- Dans le cas de la distribution normale: $C(x) = 0$, $F_1(x) = x$, $F_2(x) = x^2$, $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/2\sigma^2$, et $\psi(\theta) = \mu^2/2\sigma^2 + log(\sqrt{2\pi}\sigma)$.

Faux ami: paramètre

- En géométrie, la variété est l'objet fondamental et une paramétrisation est une construction permettant de la décrire.
- En statistiques, les paramètres ont (en général) une signification, parfois même indépendante du modèle (e.g. l'espérance).

Système de coordonnées sur les lois normales:

- Paramètres classiques: $\xi_1 = \mu$, $\xi_2 = \sigma$.
- Paramètres exponentiels (naturels): $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/2\sigma^2$.
- Paramètres en espérance: $\eta_1 = E[F_1(X)] = \mu$, $\eta_2 = E[F_2(X)] = \mu^2 + \sigma^2$.

Table des matières

Estimation

Lorsque l'on dispose d'observations $X_1, \cdots, X_n$ indépendantes et identiquement distribuées, de loi $p(., \xi) \in \mathcal{P}$, il est raisonnable de se demander quelle est la valeur de ξ.

Un estimateur $\hat{\xi}(X_1, \cdots, X_n)$ est une statistique à valeurs dans Ξ. Si l'échantillon est aléatoire, c'est une *variable aléatoire*.

En particulier, ses propriétés de convergence peuvent être étudiées.

Question: Il est naturel de regarder $\|\phi(\hat{\xi}) - \phi(\xi)\|$ comme une mesure de proximité entre l'estimateur et la vraie valeur, mais quelle fonction ϕ utiliser?

Estimation (illustration)

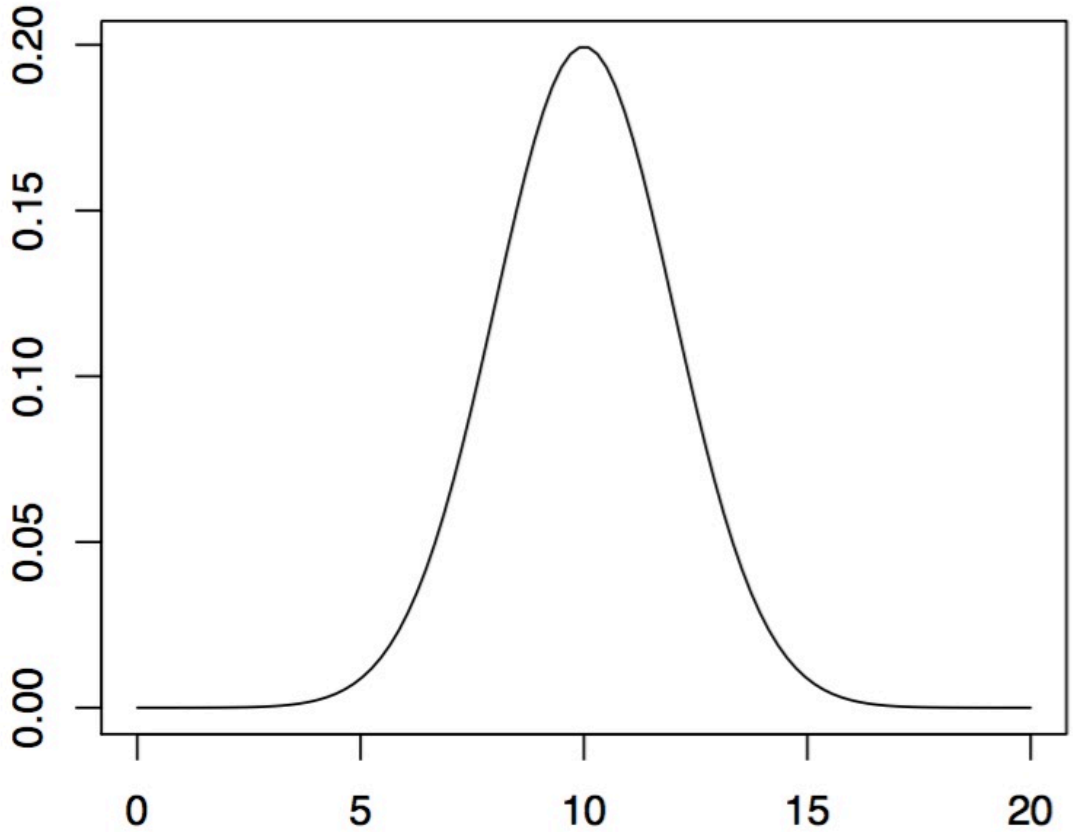


Figure: Loi sous-jacente. Paramètres fixés.

Estimation (illustration)

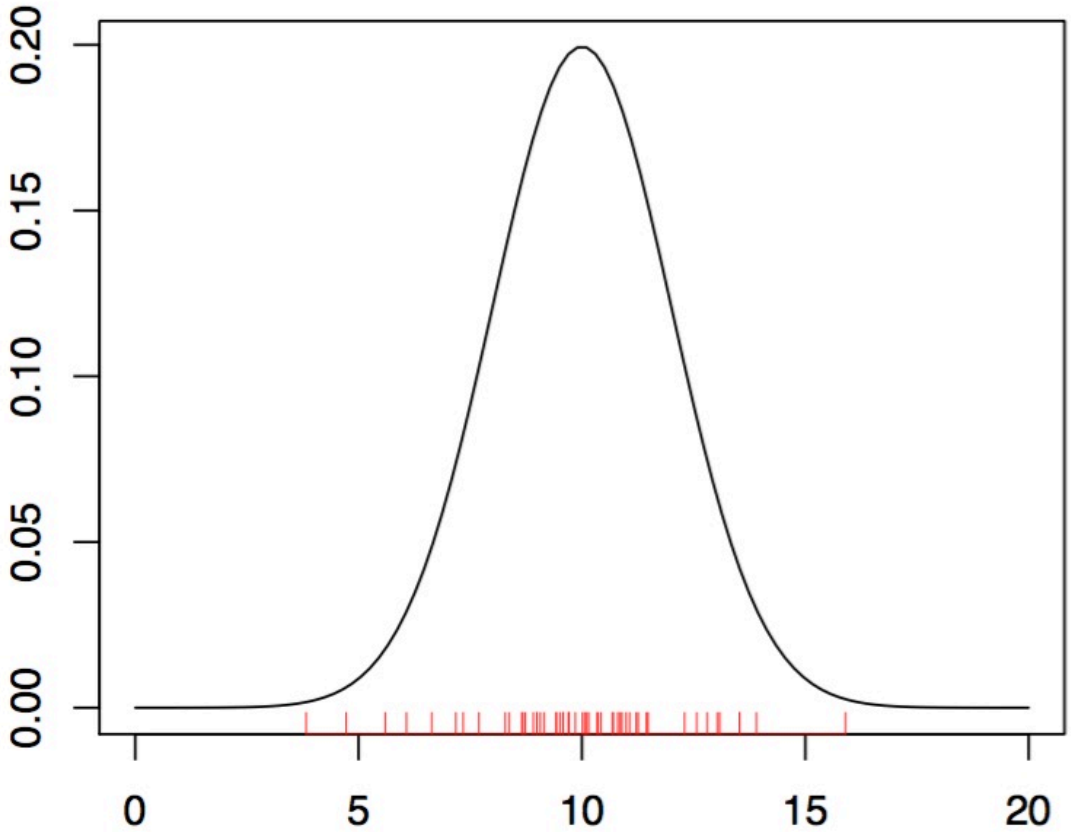


Figure: Observations engendrées.

Estimation (illustration)

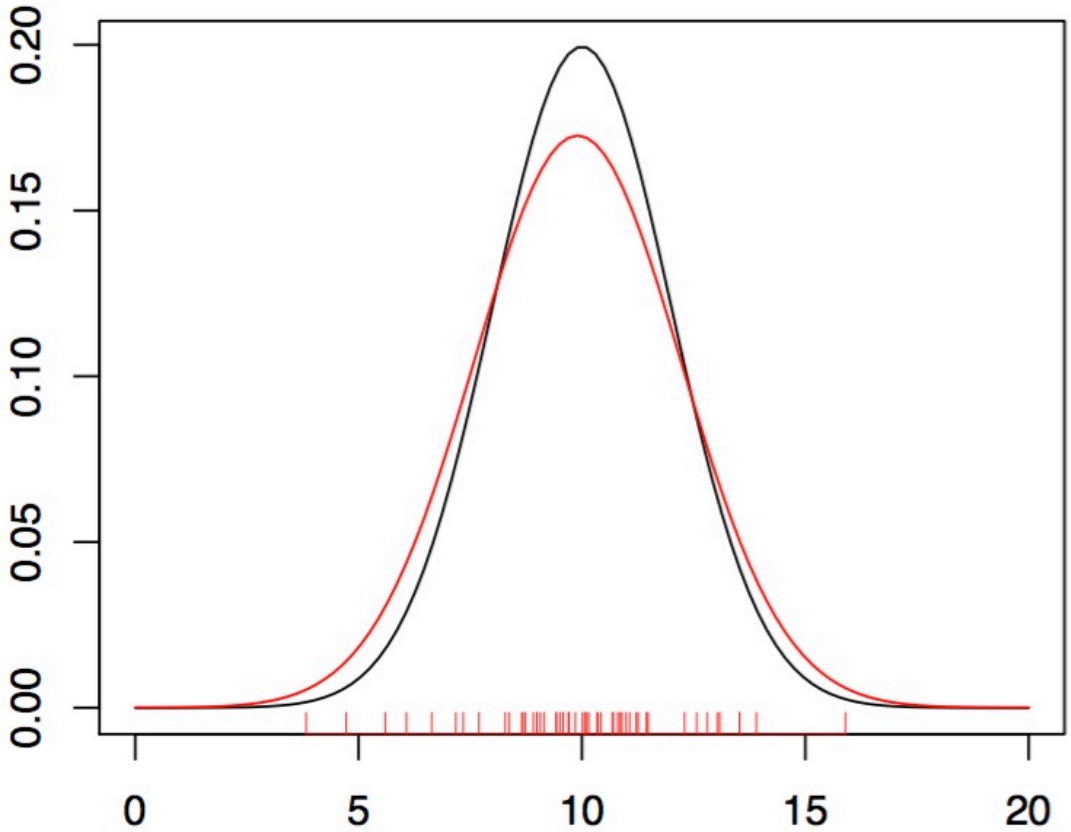


Figure: Proximité des deux courbes?

Efficacité d'un estimateur

Plusieurs propriétés sont d'importance:

- Non-biais: $E[\hat{\xi}] = \xi$,
- Convergence (faible): $\hat{\xi} \xrightarrow{P} \xi$,
- Matrice de variance-covariance: $V_\xi(\hat{\xi}) = E[(\hat{\xi} - E[\hat{\xi}])(\hat{\xi} - E[\hat{\xi}])^T]$
- Erreur quadratique moyenne: $tr\left(E[(\hat{\xi} - \xi)(\hat{\xi} - \xi)^T]\right)$
- Un estimateur sans biais est efficace si son erreur quadratique moyenne (i.e. la trace de sa matrice de covariance, i.e. $E[||e(\hat{\xi}, \xi)||^2]$) est minimale dans la classe des estimateurs sans biais.

Remarque: De manière plus générale, on pourrait s'intéresser à des estimateurs minimisant d'autres types de distance moyenne (liens avec la notion de divergence).

Table des matières

Vecteur tangent

Un vecteur tangent est un vecteur tangent à une courbe en un point.

L'espace tangent à S en p, T_pS, est l'espace vectoriel obtenu en "linéarisant" S en p.

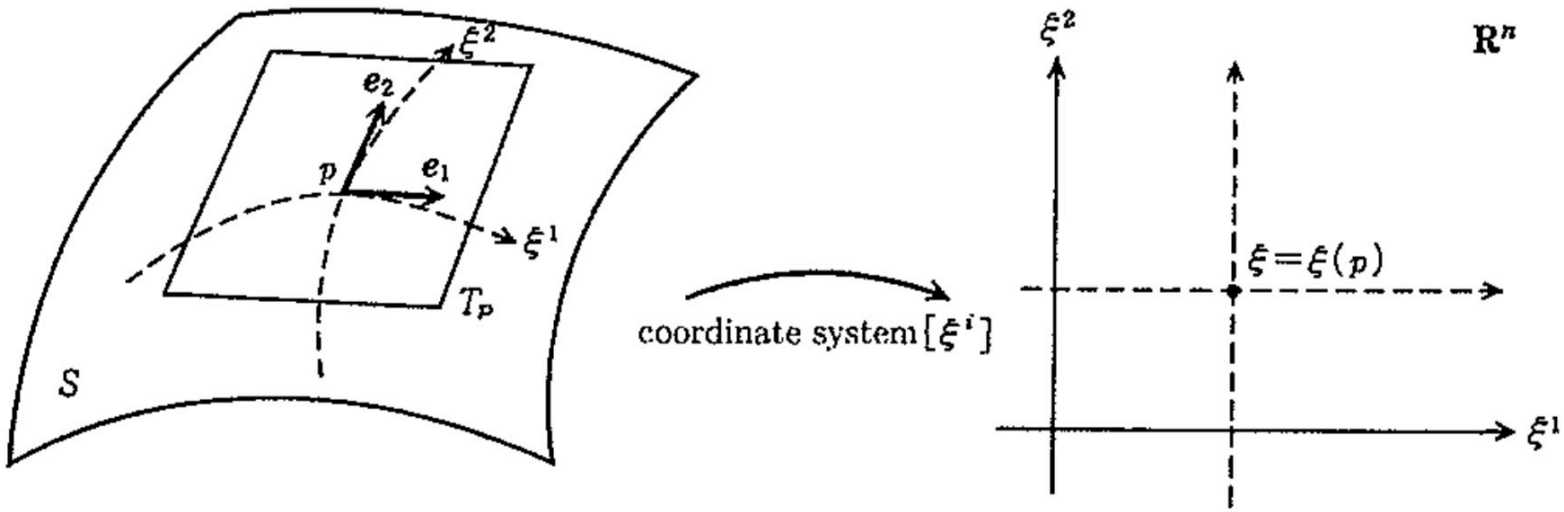


Figure: Une base de T_pS est donnée par $\{e_i, i = 1, \cdots, n\}$

Vecteur tangent (2)

- Soit $\gamma : I \to S$ une courbe dans S telle que $\gamma(a) = p$. Lorsque S est un ouvert de $\mathbb{R}^n$,
$$\dot{\gamma}(a) = \lim_{h \to 0} \frac{\gamma(a+h) - \gamma(a)}{h}.$$
- La dérivée directionnelle de $f \in C^\infty(S) = \mathcal{F}$ le long de γ est $\frac{d}{dt} f(\gamma(t))$.

Définition

Le vecteur tangent à γ en p est l'opérateur $\dot{\gamma}(a) = \sum_{i=1}^n \dot{\gamma}^i(a) \left(\frac{\partial}{\partial \xi^i} \right)_p$,

tel que $\gamma^i(t) = \xi^i \circ \gamma(t)$, $\dot{\gamma}^i(a) = \frac{d}{dt} \gamma^i(t)|_{t=a}$ *et* $(\frac{\partial}{\partial \xi^i})_p$ *est l'opérateur:* $f \mapsto (\frac{\partial f}{\partial \xi^i})_p$.

Espace tangent

Définition

L'espace tangent à S en p est la collection des vecteurs tangents S en p.

$$T_p(S) = \left\{ c^i \left(\frac{\partial}{\partial \xi^i} \right)_p \middle| [c^1, \cdots, c^n] \in \mathbb{R}^n \right\}.$$

Bien qu'écrit ci-dessus dans des coordonnées, cet espace est défini indépendamment de celles-ci! Un vecteur tangent est simplement un opérateur associant à f sa dérivée directionnelle le long d'une courbe γ.

Métrique Riemanienne

Définition

Une métrique Riemanienne est la donnée pour chaque point $p \in S$ *d'un produit scalaire* $< ., . >_p$ *défini sur* $T_p(S)$*, c'est-à-dire,* $\forall D, D' \in T_p(S)$, $< D, D' >_p \in \mathbb{R}$ *et* $< ., . >_p$ *est*

- *linéaire:* $< aD + bD', D'' >_p = a < D, D'' >_p + b < D', D'' >_p$,
- *symétrique:* $< D, D' >_p = < D', D >_p$*, et*
- *défini positif:* $< D, D >_p \geq 0$ *avec égalité ssi* $D = 0$.

Une métrique Riemanienne $g : p \mapsto < ., . >_p$ est un tenseur covariant d'ordre 2. Ses coordonnées sont $g_{ij} = < \partial_i, \partial_j >$. En particulier, $< D, D' >_p = \sum_i \sum_j D_i D'_j g_{ij}(p)$.

(S, g) est appelée une *variété Riemanienne*. Remarquons que g n'est aucunement déterminé par S.

Métrique Riemanienne (2)

Exemples:

- $X, Y \in \mathbb{R}^n$: $< X, Y >= X'MY$ pour M une matrice définie positive.
- Pour des matrices carrées: $< A, B >= tr(A^T B)$.
- Pour des variables aléatoires: $< X, Y >= E[XY]$.

Définition

Soit $\gamma : [a, b] \to S$ *une courbe sur* (S, g). *La longueur de* γ *est*

$$||\gamma|| := \int_a^b \left\| \frac{d\gamma}{dt} \right\| dt = \int_a^b \sqrt{\sum_{i,j} g_{ij} \dot{\gamma}_i \dot{\gamma}_j}.$$

Information de Fisher

L'*information de Fisher* est la métrique dont les composantes dans un système de coordonnées sont données par $G(\xi) = [g_{ij}(\xi)]$, et

$$g_{ij}(\xi) = E_\xi \left[\partial_i \ell_\xi \partial_j \ell_\xi\right] := \int \partial_i \ell(x;\xi)\partial_j \ell(x;\xi) p(x;\xi) dx,$$

où $l_\xi = l(x;\xi) = \log p(x;\xi)$ et E_ξ est l'espérance sous la loi p_ξ.

Dans le cas de la loi normale:

$$g_{ij}(\mu,\sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}.$$

Remarque: la métrique n'est plus diagonale pour les autres paramétrisations.

Statistique exhaustive

Définition

Pour une transformation $Y = F(X)$ et pour $p(x;\xi)$ la distribution de X, on a

$$p(x;\xi) = q(F(x);\xi)r(x;\xi).$$

Si $r(x;\xi)$ ne dépend pas de ξ pour tout x, la statistique $Y = F(X)$ est dite exhaustive.

Intuitivement, une statistique exhaustive est une statistique qui contient toute l'information nécessaire pour estimer le paramètre.

Théorème

Soit $G(\xi)$ l'information de Fisher de $S = p(x;\xi)$ et $G_F(\xi)$ l'information de Fisher du modèle induit $S_F = q(y;\xi)$. Alors, $G_F(\xi) \leq G(\xi)$, i.e. $G(\xi) - G_F(\xi)$ est définie positive. Cette différence est nulle si F est une statistique exhaustive.

Preuve

L'information du modèle induit est $(G_F(\xi))_{ij} = E_\xi[\partial_i \log q(Y;\xi)\partial_j \log q(Y;\xi)]$.

(1) En écrivant $p(x;\xi) = q(F(x);\xi)r(x;\xi)$, il suit

$$\partial_i \ell(x;\xi) = \partial_i \log q(F(x);\xi) + \partial_i \log r(x;\xi).$$

(2) Il est vrai que

$$\partial_i \log q(y;\xi) = E_\xi \left[\partial_i \ell(X;\xi)|y\right].$$

En effet, pour tout $B \subset \mathcal{Y}$,

$$\int_B \partial_i \log q(y;\xi) q(y;\xi) dy = \int_{F^{-1}(B)} \partial_i \ell(x;\xi) p(x;\xi) dx.$$

(3) Dès lors,

$$E_\xi \left[\partial_i \log r(X;\xi)|F(x)\right] = 0$$

Preuve (suite)

Ceci signifie que la fonction (de x) $\partial_i \log r(x;\xi)$ est orthogonale à toute fonction $\phi(F(x))$ pour le produit scalaire

$$\langle \Phi, \Psi \rangle_\xi = E_\xi[\Phi(X)\Psi(X)].$$

La perte d'information est donc donnée par

$$(\Delta G(\xi))_{ij} = E_\xi \left[\partial_i \log r(X;\xi) \partial_j \log r(X;\xi)\right] = E_\xi[Cov_\xi[\partial_i \ell(X;\xi), \partial_j \ell(X;\xi)|Y]].$$

$G_F \leq G$ s'ensuit. L'égalité sera vraie lorsque $\partial_i \log r(x;\xi) = 0$ pour tout ξ, i, x.

□

Borne de Cramér-Rao

Théorème

La variance d'un estimateur non biaisé est au moins l'inverse de l'information de Fisher:

$$V_\xi(\hat{\xi}) \geq G(\xi)^{-1}.$$

Un estimateur satisfaisant l'égalité est donc efficace. Ceci dépend crucialement de la paramétrisation!

Une condition nécessaire et suffisante pour l'existence d'un estimateur efficace sera donnée plus loin.

Preuve intuitive (1)

Soit $A : \mathcal{X} \to \mathbb{R}$ une variable aléatoire. Soit $E[A]$ la fonction $p \mapsto E_p[A]$. Le Lemme suivant montre que la variance de A (sous p) peut être vue comme la sensitivité de la fonction espérance à une perturbation de p.

Pour une fonction f, $(df)_p \in T_p^*(S)$ est définie par $(df)_p : X \mapsto X(f)$. Le gradient de f, $(\text{grad } f)_p$, est défini via

$$\langle (\text{grad } f)_p, X \rangle_p = (df)_p(X) = X(f) \quad \forall X \in T_p(S).$$

Lemme

Soit A une variable aléatoire. On a

$$V_p(A) = ||(dE[A])_p||_p^2,$$

où la norme est induite par la métrique de Fisher.

Preuve intuitive (2)

Corollaire

Pour une sous-variété $\mathcal{P} \subset \mathcal{P}_{\mathcal{X}}$,

$$V_p[A] \geq ||(dE[A]|_{\mathcal{P}})_p||_p^2.$$

La preuve de l'inégalité de Cramér-Rao suit en prenant $A = \hat{\xi}$ et en remarquant que

$$(\text{grad } f)_p = \sum_{i,j} (\partial_i f)_p (G^{-1})_{ij} (\partial_j)_p \text{ et } ||(\text{grad } f)_p||_p^2 = \sum_{i,j} (\partial_i f)_p (\partial_j f)_p (G^{-1})_{ij}$$

□

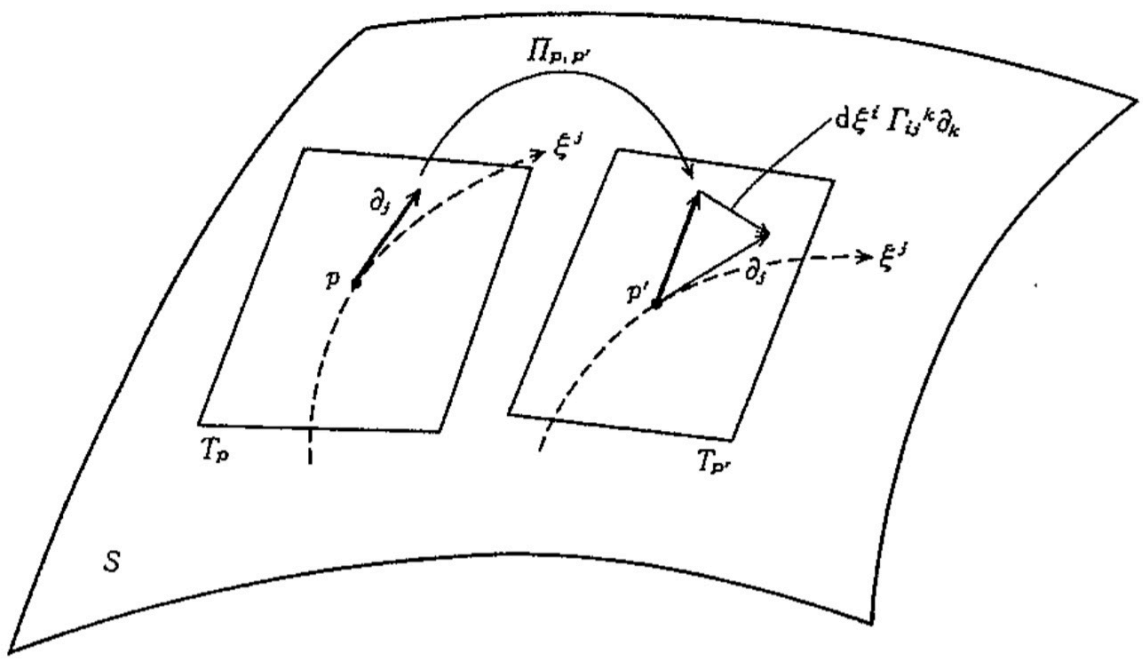
Table des matières

Connexion affine

Il reste maintenant à comprendre comment les plans tangents interagissent.

Définition (vague)

Une connexion affine est la donnée d'une transformation affine $\Pi_{p,p'}$ *entre deux plans tangents "proches".*



Symboles de Christoffel

- Notons $d\xi_i = \xi_i(p') - \xi_i(p)$.
- Si la différence entre les coordonnées de p et p' est suffisamment petite (afin d'ignorer $d\xi_i d\xi_j$), il est possible d'exprimer la différence entre $\Pi_{p,p'}\left((\partial_j)_p\right)$ et $(\partial_j)'_p$ comme une combinaison linéaire de $d\xi_1, \cdots, d\xi_n$:

$$\Pi_{p,p'}\left((\partial_j)_p\right) = (\partial_j)'_p - \left(d\xi_i(\Gamma^k_{ij})_p(\partial_k)_{p'}\right).$$

- Les n^3 quantités Γ^k_{ij} sont appelées *symboles de Christoffel*. Ce sont les coefficients de la connexion dans le système de coordonnées ξ.
- Pour une connexion sur S, ses coefficients Γ^k_{ij} dépendent des coordonnées.
- En particulier, si $\Gamma^k_{ij} = 0$, on a $\Pi_{p,p'}\left((\partial_j)_p\right) = (\partial_j)'_p$. Ces coefficients seront non-nuls dans d'autres systèmes de coordonnées.
- Exemple: Dans le plan euclidien: coordonnées canoniques vs coordonnées polaires.

Transport parallèle

Soit $\gamma : [a, b] \to S$ une courbe dans S et soit $X(t)$ un champ de vecteurs tangents le long de γ.

Définition

On dit que X est parallèle le long de γ si, pour un déplacement infinitésimal dt, il existe une transformation linéaire $\Pi_{p,p'}$ ($p = \gamma(t)$, $p' = \gamma(t + dt)$) telle que

$$X(t + dt) = \Pi_{p,p'}(X(t)).$$

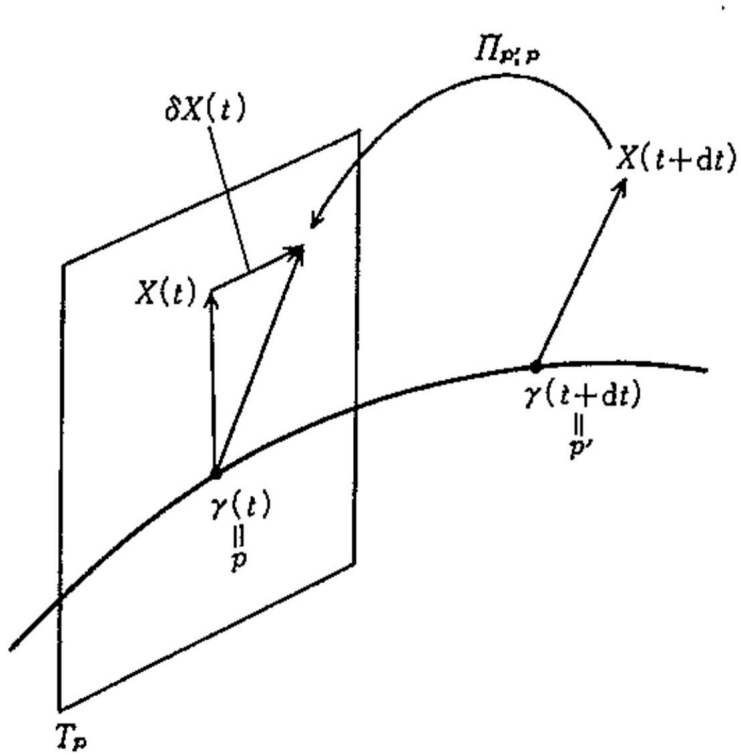
- Un champ de vecteurs peut-être parallèle pour une connexion mais pas pour une autre.
- Ceci permet également de définir Π_γ, le transport parallèle le long de γ, par recollement.

Dérivée covariante le long d'une courbe

Soit $X(t)$ in champ de vecteurs le long de γ. Comme $X(t+h)$ et $X(t)$ sont dans des espaces tangents distincts, l'expression

$$\frac{dX(t)}{dt} = \lim_{h \to 0} \frac{X(t+h) - X(t)}{h}$$

ne fait, en général, pas de sens. La donnée d'une connexion permet cependant de définir une "dérivée".



Dérivée covariante le long d'une courbe (2)

Définition

La dérivée covariante de X le long de γ en t est

$$\frac{\delta X}{dt} = \lim_{h \to 0} \frac{X_t(t+h) - X(t)}{h} := \lim_{h \to 0} \frac{\Pi_{\gamma(t+h),\gamma(t)}(X(t+h)) - X(t)}{h}$$

- En particulier, X est parallèle le long de γ si $\frac{\delta X}{dt} = 0$.
- La dérivée directionnelle d'un champ de vecteurs X dans S dans la direction $D \in T_p(S)$ est obtenue en prenant la dérivée covariante le long de toute courbe de vecteur tangent D en p. Elle est notée $\nabla_D X$. Plus généralement, $\delta X_{\gamma(t)}/dt = \nabla_{\dot{\gamma}(t)} X$.

Dérivée covariante de champs de vecteurs

Définition

Soient X et Y deux champs de vecteurs sur S. La dérivée covariante de Y le long de X, $\nabla_X Y$ est le champ de vecteurs défini par

$$(\nabla_X Y)_p = \nabla_{X_p} Y.$$

En particulier, on a $\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k \Rightarrow$ interprétation de Γ_{ij}^k.

L'opérateur $\nabla : \mathcal{T}(S) \times \mathcal{T}(S) \to \mathcal{T}(S)$ définit la connexion (sous certaines conditions). On parlera dans la suite de la *connexion* ∇.

Reformulation équivalente: $\Gamma_{ij,k} := \langle \nabla_{\partial_i} \partial_j, \partial_k \rangle = \sum_h \Gamma_{ij}^h g_{hk}$.

Connexion métrique et connexion de Levi-Civita

Jusqu'ici, le lien entre métrique et connexion est inexistant.

Définition

Si, pour tout champs de vecteurs $X, Y, Z \in \mathcal{T}(S)$,

$$Z \langle X, Y \rangle = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle ,$$

où $Z \langle X, Y \rangle$ *est la dérivée du produit scalaire le long de* Z*, on dit que* ∇ *est une connexion métrique (de métrique* g*).*

- La transport parallèle via une connexion métrique préserve le produit scalaire (i.e. est une isométrie):
$$\langle \Pi_\gamma(D_1), \Pi_\gamma(D_2) \rangle_q = \langle D_1, D_2 \rangle_p .$$
- La connexion de Levi-Civita est l'unique connexion métrique et symétrique ($\Gamma^k_{ij} = \Gamma^k_{ji}$). Elle est aussi appelée *connexion Riemanienne*.
- En coordonnées, la condition s'écrit $\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}$.

α-connexions

Soit S un modèle paramétrique. Soient les fonctions

$$\xi \mapsto (\Gamma_{ij,k})_\xi = E_\xi \left[\left(\partial_i \partial_j \ell_\xi + \frac{1}{2} \partial_i \ell_\xi \partial_j \ell \xi \right) (\partial_k \ell_\xi) \right].$$

Théorème

Les coefficients $\Gamma_{ij,k}$ sont les coefficients de la connexion Riemanienne associée à la métrique de Fisher.

Preuve: En dérivant les composantes de la métrique de Fisher, on trouve

$$\begin{aligned} \partial_k g_{ij} &= E_\xi \left[(\partial_k \partial_i \ell_\xi)(\partial_j \ell_\xi) \right] + E_\xi \left[(\partial_i \ell_x i)(\partial_k \partial_j \ell_\xi) \right] + E_\xi \left[(\partial_i \ell_\xi)(\partial_j \ell_\xi)(\partial_k \ell_\xi) \right] \\ &= \Gamma_{ki,j} + \Gamma_{kj,i}. \end{aligned}$$

□

α-connexions (2)

Il sera intéressant par la suite de définir la famille de connexions suivante:

Définition

Pour un modèle paramétrique $S = \{p_\xi\}$, la connexion α, $\nabla^{(\alpha)}$ est la connexion satisfaisant $< \nabla^{(\alpha)}_{\partial_i} \partial_j, \partial_k >= \Gamma^{(\alpha)}_{ij,k}$, où

$$\left(\Gamma^{(\alpha)}_{ij,k}\right)_\xi = E_\xi \left[\left(\partial_i \partial_j \ell_\xi + \frac{1-\alpha}{2} \partial_i \ell_\xi \partial_j \ell \xi \right) (\partial_k \ell_\xi) \right].$$

- $\nabla^{(0)}$ est la connexion Riemanienne "de Fisher".
- Par additivité des connexions, $\nabla^{(\alpha)} = \frac{1+\alpha}{2} \nabla^{(1)} + \frac{1-\alpha}{2} \nabla^{(-1)}$.
- $\nabla^{(\alpha)}$ est symétrique pour tout α.

Théorème de Chentsov

En quoi la métrique de Fisher et ces connexions sont-elles uniques?

Remarquons que, pour un modèle S, $F(X)$ une statistique exhaustive et S_F le modèle induit, on a $\partial_i \log p(x;\xi) = \partial_i \log q(F(x);\xi)$ et ainsi g_{ij} et $\Gamma^{(\alpha)}_{ij,k}$ coincident.

Théorème (Chentsov (1972) & Ay *et al.* (2015))

Supposons (g, ∇) invariant pour les statistiques exhaustives. Alors, il existe une constante c et un nombre réel α tels que cg est la métrique de Fisher et $\nabla = \nabla^{(\alpha)}$.

Coordonnées affines

- Si, pour toute courbe γ sur S, un champ de vecteurs X est parallèle le long de γ, on dit que X est parallèle sur S.
- Dans ce cas, $X_q = \Pi_\gamma(X_p)$ pour tout γ ou encore $\nabla_Y X = 0$ pour tout $Y \in \mathcal{T}(S)$.

Définition

Un système de coordonnées affines *sur S (pour la connexion ∇) est un système ξ tel que $\partial_i = \partial/\partial\xi_i$ est parallèle sur S.*

Des conditions équivalentes sont $\nabla_{\partial_i}\partial_j = 0$ ou encore $\Gamma^k_{ij} = 0$.

Planéité (flatness)

Pour une connexion ∇ donnée, il n'existe pas nécessairement un système affin associé.

Définition

S est plat pour la connexion ∇ si un système affin existe sur S pour ∇.

Théorème

Soient ξ et ρ deux systèmes affins sur S. Alors $\xi(p) = A\rho(p) + b$, pour A une matrice $n \times n$ et $b \in \mathbb{R}^n$.

Courbure et torsion

Définition

La courbure $R : \mathcal{T} \times \mathcal{T} \times \mathcal{T} \to \mathcal{T}$ *d'une connexion* ∇ *sur* S *est caractérisée par, pour* $X, Y, Z \in \mathcal{T}(S)$,

$$R(X,Y)Z = \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X,Y]}Z.$$

Définition

La torsion $R : \mathcal{T} \times \mathcal{T} \to \mathcal{T}$ *d'une connexion* ∇ *sur* S *est caractérisée par, pour* $X, Y \in \mathcal{T}(S)$,

$$T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y].$$

- En composantes: $R(\partial_i, \partial_j)\partial_k = R^l_{ijk}\partial_l$ et $T(\partial_i, \partial_j) = T^k_{ij}\partial_k$, où
$$R^l_{ijk} = \partial_i \Gamma^l_{jk} - \partial_j \Gamma^l_{ik} + \Gamma^l_{ih}\Gamma^h_{jk} - \Gamma^l_{jh}\Gamma^h_{ik} \text{ et } T^k_{ij} = \Gamma^k_{ij} - \Gamma^k_{ji}.$$
- La courbure est nulle si les transports parallèles ne dépendent pas de la courbe choisie.

Géodésiques

Définition

Une géodésique, ou courbe autoparallèle, est une sous-variété autoparallèle de dimension 1.

- Ceci est une propriété liée à la connexion ∇, non la métrique.
- Si la connexion est la connexion Riemanienne, on peut montrer qu'une géodésique est la courbe de longueur minimale entre deux points.
- Si, de plus, S est plat pour ∇, alors les géodésiques sont les préimages des droites dans le système de coordonnées.

Courbure de la famille normale

La famille des lois normales est courbée, de courbure constante. Sa courbure scalaire sous $\nabla^{(\alpha)}$ est donnée par $-(1-\alpha)^2/2$. En particulier, pour $\alpha = 0$, sa courbure est négative et l'espace métré des paramètres est la *plan de Poincaré*.

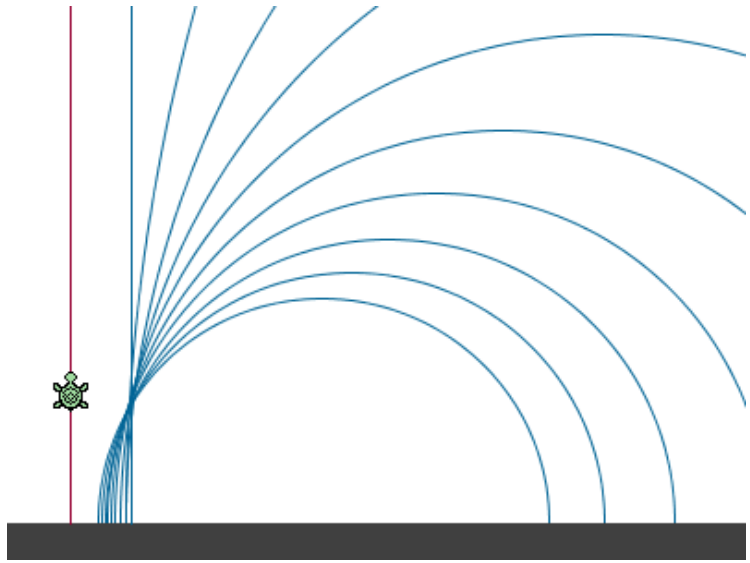


Figure: Des droites parallèles dans le plan de Poincaré

Cette courbure constante sous les connexions α est une caractérisation des familles position-échelle.

Planéité de modèles statistiques (famille exponentielle)

Revenons sur une famille exponentielle: $p(x;\theta) = exp\left(C(x) + \sum_{i=1}^{n} \theta_i F_i(x) - \psi(\theta)\right)$.
Il est facile de voir que, pour $\partial_i = \partial/\partial\theta_i$,

$$\partial_i \ell(x;\theta) = F_i(x) - \partial_i \psi(\theta) \text{ et } \partial_i \partial_j \ell(x;\theta) = -\partial_i \partial_j \psi(\theta).$$

En particulier,

$$\Gamma^{(1)}_{ij,k} = -\partial_i \partial_j \psi(\theta) E_\theta[\partial_k \ell_\theta] = 0.$$

Théorème

Dans une famille exponentielle, le système de coordonnées θ est affin pour la connexion $\nabla^{(1)}$. En particulier, le modèle est 1-plat.

La connexion $\nabla^{(1)} = \nabla^{(e)}$ est souvent appelée *connexion exponentielle*.

Planéité de modèles statistiques (mélanges)

Soit le modèle $S = \{p_\theta\}$ (modèle de mélange) tel que

$$p(x;\theta) = C(x) + \sum_{i=1}^{n} \theta_i F_i(x).$$

Il est facile de voir que

$$\partial_i \ell(x;\theta) = \frac{F_i(x)}{p(x;\theta)} \text{ et } \partial_i \partial_j \ell(x;\theta) = -\frac{F_i(x)F_j(x)}{p(x;\theta)^2}.$$

En particulier,

$$\Gamma^{(-1)}_{ij,k} = E_\theta[(\partial_i \partial_j \ell_\theta + \partial_i \ell_\theta \partial_j \ell_\theta)(\partial_k \ell_\theta)] = 0.$$

Théorème

Dans un modèle de mélange, le système de coordonnées θ est affin pour la connexion $\nabla^{(-1)}$. En particulier, le modèle est (-1)-plat.

La connexion $\nabla^{(-1)} := \nabla^{(m)}$ est souvent appelée *connexion de mélange*.

Table des matières

Dualité

Définition

Soit S une variété munie d'une métrique g et de deux connexions ∇ et ∇^*. Ces deux connexions sont duales *si, pour $X, Y, Z \in \mathcal{T}(S)$,*

$$Z < X, Y >=< \nabla_Z X, Y > + < X, \nabla^*_Z Y > .$$

Le triple (g, ∇, ∇^*) est appelée structure duale *sur S.*

- Une connexion ∇ est métrique si $\nabla = \nabla^*$.
- $(\nabla^*)^* = \nabla$.
- $\left\langle \Pi_\gamma(D_1), \Pi^*_\gamma(D_2) \right\rangle_q = \langle D_1, D_2 \rangle_p$.
- $R = 0 \Leftrightarrow R^* = 0$.
- Soit (g, ∇, ∇^*) une structure duale sur S. Si les connexions ∇ et ∇^* sont symétriques, alors S est plat pour ∇ ssi S est plat pour ∇^*.

Définition

(S, g, ∇, ∇^*) est un espace dualement plat si ∇ et ∇^* sont plats.

Dualité (stat)

Pour tout modèle statistique, les connexions $\nabla^{(\alpha)}$ et $\nabla^{(-\alpha)}$ sont duales pour la métrique de Fisher.

Puisque les connexions $\nabla^{(\alpha)}$ sont symétriques,

Théorème

S est α-plat $\Leftrightarrow$ S est $(-\alpha)$-plat.

En particulier, les familles exponentielles et de mélange sont (± 1)-plates.

Systèmes duaux de coordonnées

Soit θ un système de coordonnées ∇-affin. Si η est un système de coordonnées ∇^*-affin tel que

$$\left\langle \partial_{\theta_i}, \partial_{\eta_j} \right\rangle = \delta_{ij},$$

ces deux systèmes sont appelés (mutuellement) duaux pour la métrique g.

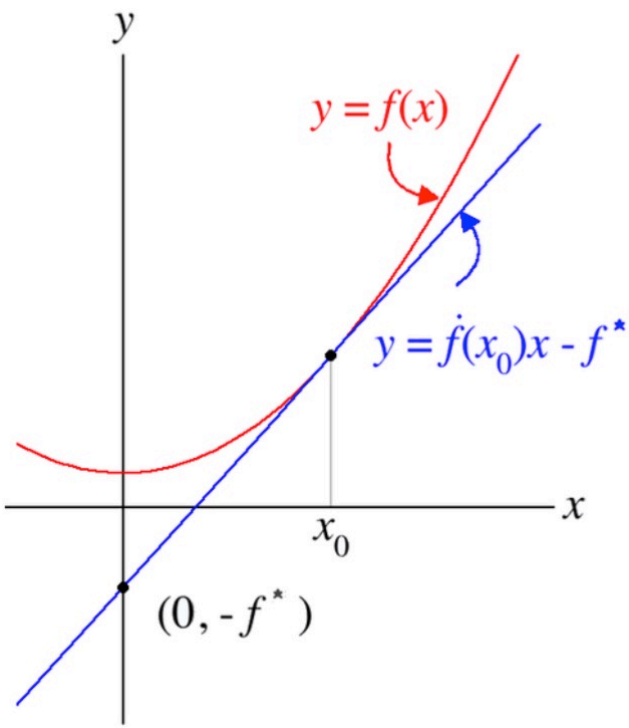
Théorème

Une paire de systèmes duaux de coordonnées existe si et seulement si (S, g, ∇, ∇^*) est dualement plat.

Transformation de Legendre

Soient θ et η deux systèmes de coordonnées mutuellement duaux. Soient les fonctions $\psi : S \to R$ et $\phi : S \to R$ telles que (i) $\partial_i \psi = \eta_i$, (ii) $\partial_i \phi = \theta_i$, (iii) $g_{ij} = \partial_i \eta_j = \partial_j \eta_i = \partial_i \partial_j \psi$, (iv) $\phi(\eta) = \max_\theta \{\theta_i \eta_i - \psi(\theta)\}$ et (v) $\psi(\theta) = \max_\eta \{\theta_i \eta_i - \phi(\eta)\}$

Ces deux fonctions ψ et ϕ (i) existent et (ii) permettent le passage d'un système de coordonnées à l'autre via la *transformation de Legendre*. Elles sont appelées *potentiels* de θ et η.



Interprétation géométrique de $f^*(p) = \max_x(xp - f(x))$, $p = f'(x)$:
Pour une fonction convexe f, pour $p_0 = f'(x_0)$ fixé, la droite tangente en $(x_0, f(x_0))$ intersecte l'axe des ordonnées en un point d'ordonnée $-f^*(x_0)$. C'est en effet le maximum de l'expression ci-dessus, à p fixé.

Transformation de Legendre

Soient θ les coordonnées naturelles d'une famille exponentielle. Les coordonnées en espérance η, définies par $\eta_i = E_\theta[F_i(X)]$ est le système dual de θ. Il est (-1)-affin.

Dans le cadre de la loi normale, on a $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/2\sigma^2$ et

$$\psi(\theta) = \frac{-(\theta_1)^2}{4\theta^2} + \frac{1}{2}\log(\frac{\pi}{\theta^2}) = \frac{\mu}{2\sigma^2} + \log(\sqrt{2\pi}\sigma).$$

Les paramètres duaux sont $\eta_1 = \partial\psi/\partial\theta_1 = \mu = -\theta_1/2\theta_2$ et $\eta_2 = \partial\psi/\partial\theta_2 = \mu^2 + \sigma^2$. La fonction potentielle associée est

$$\phi(\eta) = -\frac{1}{2}\left(1 + \log(2\pi) + 2\log(\sigma)\right).$$

Table des matières

Divergence

Définition

Soit S une variété. Une divergence sur S est une fonction $D = D(.||.) : S \times S \to R$, C^∞, satisfaisant pour tout $p, q \in S$

$$D(p||q) \geq 0,$$

avec égalité ssi $p = q$.

Remarque: Ce n'est pas ni une semi-distance (+symétrie), ni une distance (+inégalité triangulaire).

Kullback-Leibler

Définition

Pour deux lois continues p et q, la divergence de Kullback-Leibler est définie par

$$D_{KL}(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx.$$

Cette divergence mesure la différence entre deux distributions. En général, p représente la "vraie" loi et q est la loi estimée. Cette divergence n'est pas symétrique et ne satisfait pas l'inégalité triangulaire.

Bregman

Définition

Soit Δ un ensemble convexe fermé de $\mathbb{R}^n$. La divergence de Bregman associée à la fonction convexe F sur Δ est

$$D_B(x||y) = F(y) - F(x) - \langle y - x, \nabla F(x) \rangle$$

Pour $F(x) = ||x||^2$, on a $D_B = ||x - y||^2$. Pour $F(x) = x \log(x) - x$, on retrouve la divergence de KL.

Divergences statistiques

La notion de divergence est utilisée abondamment dans le contexte de l'estimation:

Définition

Soit D une divergence. L'estimateur minimum de divergence est la valeur $\hat{\xi}$ telle que

$$D(\hat{p}||p_{\hat{\xi}}) = \min_{\xi} D(\hat{p}||p_{\xi}).$$

- Equivalent à un estimateur maximum de quasi-vraisemblance.
- Permet la définition d'estimateurs robustes, asymptotiquement efficaces, etc. Champ de recherche actif.

Métrique et connexions induites

Soit D une divergence. Il est alors possible de définir une métrique sur S via

$$\langle X, Y \rangle^{(D)} = -D[X||Y] := -\sum_{i,j} X_i Y_j \partial_i \partial'_j D[p||p'].$$

De même, une connexion affine peut être introduite:

$$\left\langle \nabla_X^{(D)} Y, Z \right\rangle^{(D)} = -D[XY||Z].$$

La divergence duale de D est $D^*(p||q) = D(q||p)$.

Théorème

$\nabla^{(D)}$ et $\nabla^{(D^*)}$ sont duales pour $g^{(D)}$.

Réciproquement, tout (g, ∇, ∇^*) mutuellement duaux sont induits par une divergence (non unique).

Divergence canonique

Définition

Soit (S, g, ∇, ∇^*) *un système dualement plat. Soit* θ *et* η *les systèmes de coordonnées affines duaux, de potentiels* ψ *et* ϕ. *La* divergence canonique (g, ∇) *est*

$$D(p||q) = \psi(p) + \phi(q) - \langle \theta(p), \eta(q) \rangle .$$

- On a $D^*(q||p) = D(p||q)$, pour D^* la divergence canonique (g, ∇^*).
- Dans le cas d'une connexion métrique, le système de coordonnées est autodual et $D(p||q) = 1/2 d(p, q)^2$.

Théorème

Soit (S, g, ∇, ∇^*) *un système dualement plat de coordonnées affines duales* θ *et* η *et soit* D *une divergence sur* S. *Une condition nécessaire et suffisante pour que* D *soit la divergence canonique* (g, ∇) *est, pour* $p, q, r \in S$

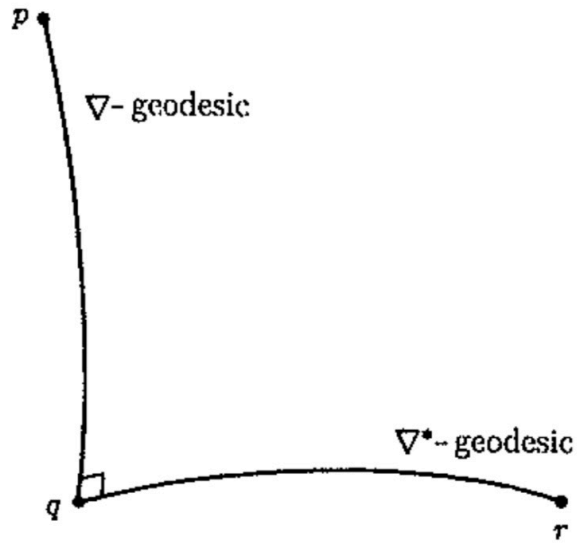
$$D(p||q) + D(q||r) - D(p||r) = \langle \theta(p) - \theta(q), \eta(p) - \eta(q) \rangle .$$

Relation de Pythagore

Théorème (Relation de Pythagore)

Soient $p, q, r \in S$. Soit γ_1 la ∇-géodésique de p à q et soit γ_2 la ∇^*-géodésique de q à r. Si les deux géodésiques sont orthogonales (au sens de la métrique g) en q, alors

$$D(p||r) = D(p||q) + D(q||r).$$



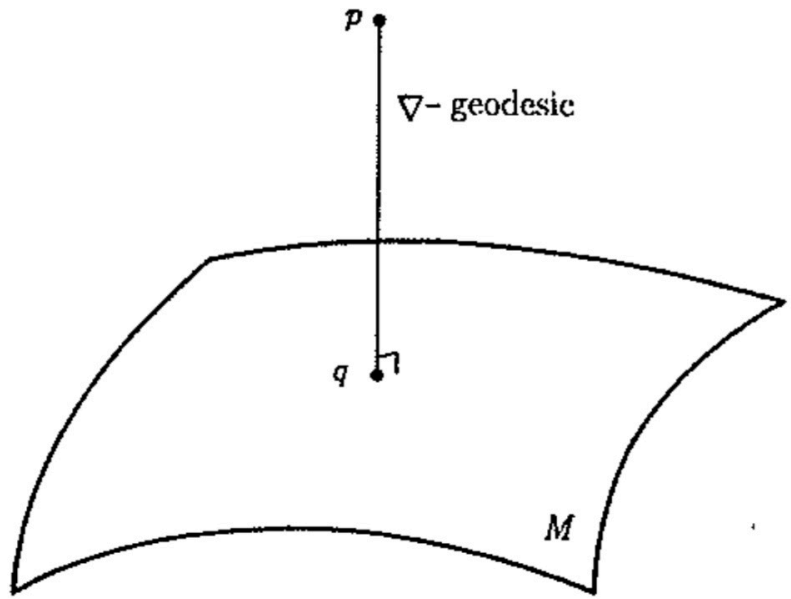
Projection

Théorème

Soit $p \in S$ et soit M une sous-variété de S qui est ∇^*-autoparallèle. Une condition nécessaire et suffisante telle que $q \in M$ satisfait

$$D(p||q) = \min_{r^{(n)}\, M} D(p||r)$$

est que la ∇-géodésique reliant p à q soit orthogonale à M en q.



Projection

L'utilité du théorème précédent est la suivante:

L'ensemble des produits de distributions

$$E = \left\{ p_X \middle| p_X(x_1, \cdots, x_N) = \prod p_{X_i}(x_i) \right\}$$

est 1-plat. Obtenir un estimateur minimum de divergence requiert donc de suivre les m-géodésiques projetant sur E.

Il permet également de prouver le résultat suivant:

Théorème

Une condition nécessaire et suffisante pour qu'un système de coordonnées possède un estimateur efficace est que S soit une famille exponentielle et que ξ soit (-1)-affin.

Table des matières

Possibles extensions

- Souvent, il est plus naturel de regarder un ensemble de probabilités comme un ensemble fermé. Il est alors nécessaire d'introduire la notion de *variétés à bord(s)*.
- Très souvent, la torsion des variétés statistiques est nulle. L'introduction d'une torsion non-nulle est utile dans les domaines de la mécanique quantique (proba non commutative), théorie des systèmes, etc.
- Les connexions α sont abondamment utilisées en mécanique statistique et sont liées à l'entropie de Tsallis.
- Les outils de géométrie algébrique permettent l'étude de modèles statistiques sur des graphes et des réseaux neuronaux.
- Enormément d'applications pratiques: imagerie médicale, transport optimal, optimisation sur variétés, etc.

Modèles courbes

Théorème

Soit S une famille exponentielle (resp. de mélange) et M une sous-variété de S. Alors, M est une famille exponentielle (resp. de mélange) si M est 1-autoparallèle (resp. (-1)).

De manière générale, une sous-variété est appelée famille courbe. L'étude de la courbure de M dans S est une des pierres angulaires de l'inférence statistique via la Géométrie de l'Information.

Merci pour votre attention

Références

Amari, S.-I. (1985). Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics, Vol. 28, Springer Verlag. Berlin, Heidelberg.

Amari, S.-I. & Nagaoka, H. (1993). Methods of Information Geometry. Translation of Mathematical Monographs, Oxford University Press.

Efron, B. (1975). Defining the curvature of a statistical problem (with discussion). *Ann.Statist.*, **3**, 1189-1242.

Rao, C.R.(1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, **37**, 81-91.