Introduction à la statistique non paramétrique

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry

http://stat.genopole.cnrs.fr/∼cmatias

Atelier SFDS
27/28 septembre 2012

Partie 1 : Introduction

Plan partie 1

Quelques références bibliographiques pour cet atelier

L. Wasserman.
All of nonparametric statistics.
Springer Texts in Statistics. Springer-Verlag, 2006.

E.L. Lehmann.
Elements of large sample theory.
Springer Texts in Statistics. Springer-Verlag, 1999.

A. B. Tsybakov.
Introduction à l'estimation non-paramétrique, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*.
Springer-Verlag, Berlin, 2004.

D. Bosq.
Nonparametric statistics for stochastic processes,
Springer-Verlag, 1996.

Statistique non paramétrique : c'est quoi ?

La statistique paramétrique est le cadre "classique" de la statistique. Le modèle statistique y est décrit par un nombre fini de paramètres. Typiquement $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \mathbb{R}^p\}$ est le modèle statistique qui décrit la distribution des variables aléatoires observées.

Exemples

- Observations réelles avec un seul mode : $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{+\star}\}$, modèle Gaussien.
- Observations réelles avec plusieurs modes : $\mathcal{M}_K = \{\sum_{i=1}^K p_i \mathcal{N}(\mu_i, \sigma^2), (p_1, \ldots, p_K) \in (0,1)^K, \sum_i p_i = 1, (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K, \sigma^2 \in \mathbb{R}^{+\star}\}$, modèle de mélange Gaussien.
- Observations de comptage : $\mathcal{M} = \{\mathcal{P}(\lambda); \lambda \in \mathbb{R}^{+\star}\}$, modèle loi Poisson.
- $\mathcal{M} = \{\mathbb{P}$ à support dans $\mathcal{S}$ fini$\} \simeq [0,1]^{|\mathcal{S}|-1}$.

Statistique non paramétrique : c'est quoi ?

Par opposition, en statistique non paramétrique, le modèle n'est pas décrit par un nombre fini de paramètres.
Divers cas de figures peuvent se présenter, comme par exemple :

- On s'autorise toutes les distributions possibles, *i.e.* on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires.
- On travaille sur des espaces fonctionnels, de dimension infinie. Exemple : les densités continues sur $[0, 1]$, ou les densités monotones sur $\mathbb{R}$.
- Le nombre de paramètres du modèle n'est pas fixé et varie (augmente) avec le nombre d'observations.
- Le support de la distribution est discret mais varie (augmente) avec le nombre d'observations.

Motivations I

Observations rangées

Situation :

- On dispose des résultats de questionnaires où des échantillons de consommateurs ont classé un ensemble de produits par ordre de préférence,
- Les questionnaires proviennent de supermarchés situés dans des zones socio-économiques différentes,
- On se demande si un produit P obtient un classement significativement différent d'un supermarché à l'autre.

Questions :

- Comment modéliser la distribution des observations ?
- Quel test utiliser ?

Réponse : Tests de rang.

Motivations II

Observations mesurées

Situation : On observe des données quantitatives.

Questions :

- Peut-on raisonnablement supposer que les observations suivent une loi normale ? (par exemple pour faire des tests sur la moyenne). Rép : Tests de normalité.
- Combien de modes possède cette distribution ? Rép : Estimation de densité.

Statistique non paramétrique : Quand l'utiliser ?

Exemples de contextes d'utilisation

- Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique,
- Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle,
- Quand on ne sait pas combien de composantes on veut mettre dans un mélange,
- Quand le nombre de variables est trop grand (problème de grande dimension) et qu'un modèle paramétrique est non utilisable car il aurait de toutes façons trop de paramètres,
- ...

Avantages/Inconvénients

Avantages

- Moins d'a priori sur les observations,
- Modèles plus généraux, donc plus robustes au modèle.

Inconvénients

- Vitesses de convergence plus lentes = il faut plus de données pour obtenir une précision équivalente.

Plan partie 1

Estimer une fonction de répartition

On observe $X_1, \ldots, X_n$ variables aléatoires (v.a.) réelles, i.i.d. de fonction de répartition (fdr) $F : x \to F(x) = \mathbb{P}(X_1 \leq x)$.
L'estimateur naturel de la fdr F est la fdr empirique $\hat{F}_n$ définie par $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}$. C'est un estimateur non paramétrique de la fdr F.

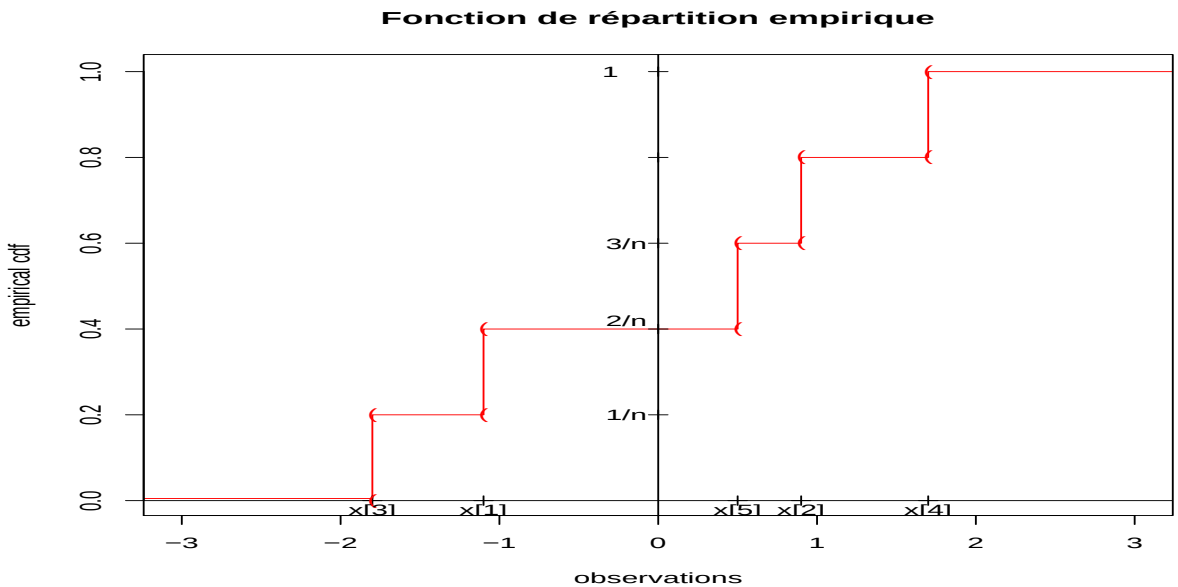


Figure : Fonction de répartition empirique.

→ Qualité de cet estimateur ?

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* x fixé) I

- Biais

$$\mathbb{E}\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i \leq x) = F(x),$$

i.e. estimateur sans biais.

- Variance

$$\mathbb{V}\mathrm{ar}(\hat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(1_{X_i \leq x}) = \frac{1}{n}\mathbb{V}\mathrm{ar}(1_{X_1 \leq x})$$

$$= \frac{F(x)(1-F(x))}{n} \xrightarrow[n\to\infty]{} 0.$$

- Erreur en moyenne quadratique (ou MSE pour "mean square error")

$$\mathbb{E}[(\hat{F}_n(x) - F(x))^2] = \text{biais}^2 + \text{variance} = \mathbb{V}\mathrm{ar}(\hat{F}_n(x)) \xrightarrow[n\to\infty]{} 0.$$

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* x fixé) II

- Convergence en probabilité

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{\text{proba}} F(x).$$

En effet, d'après l'inégalité de Markov, la convergence en moyenne quadratique implique la convergence en probabilité.

$$\forall \epsilon > 0, \quad \mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \frac{\mathbb{V}\text{ar}(\hat{F}_n(x))}{\epsilon^2} \xrightarrow[n\to\infty]{} 0.$$

- LGN :

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{\text{p.s.}} F(x).$$

- TCL :

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, F(x)(1 - F(x))).$$

Propriétés ponctuelles de $\hat{F}_n(x)$ (*i.e.* x fixé) III

- Loi du logarithme itéré LIL. Rappel : si $\{X_i\}_{i\geq 0}$ suite de v.a. i.i.d., centrées, de variance $\sigma^2 < +\infty$ et $S_n = \sum_{i=1}^n X_i$. Alors

$$\limsup_{n\to\infty} \frac{|S_n|}{\sigma\sqrt{2n\log\log n}} = 1 \quad \text{p.s.}$$

  En particulier

$$\limsup_{n\to\infty} \frac{\sqrt{n}|\hat{F}_n(x) - F(x)|}{\sqrt{F(x)(1-F(x))2\log\log n}} = 1 \quad \text{p.s.}$$

Propriétés uniformes de $\hat{F}_n$

- Théorème de Glivenko Cantelli

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

- Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW)

$$\forall n \in \mathbb{N}, \forall \epsilon > 0, \quad \mathbb{P}(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Exemple d'application de l'inégalité de DKW I

Construction d'intervalles de confiance (IC) exacts sur $F(x)$

En effet, $\forall x \in \mathbb{R}$, on a

$$\begin{aligned}\mathbb{P}(F(x) \in [\hat{F}_n(x) - \epsilon; \hat{F}_n(x) + \epsilon]) &= 1 - \mathbb{P}(|\hat{F}_n(x) - F(x)| > \epsilon) \\ &\geq 1 - \mathbb{P}(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon) \geq 1 - 2e^{-2n\epsilon^2}.\end{aligned}$$

Pour tout $\alpha > 0$, on choisit alors $\epsilon > 0$ tel que $2e^{-2n\epsilon^2} = \alpha$, *i.e.* on prend $\epsilon = \sqrt{\log(2/\alpha)/(2n)}$ et on obtient

$$\begin{aligned}\mathbb{P}(F(x) \in [\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)}; \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)}]) \\ \geq 1 - \alpha,\end{aligned}$$

donc $[\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)}; \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)}]$ est un IC au niveau $1 - \alpha$ pour $F(x)$.

Exemple d'application de l'inégalité de DKW II

Remarques

- Comme $F(x) \in [0,1]$, si n est petit on peut souvent raffiner cet IC en prenant plutôt $[\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)}; \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)}] \cap [0,1]$.
- Le TCL permet également d'obtenir un IC pour $F(x)$, à condition d'estimer la variance $F(x)(1-F(x))$. Mais cet intervalle est asymptotique uniquement. Il peut s'avérer meilleur que l'intervalle exact ci-dessus car ce dernier est fondé sur une borne uniforme qui peut être mauvaise pour certaines valeurs de x.

Autres exemples autour de la FDR empirique

Dans la suite, nous nous intéresserons également au cas des fonctionnelles de la distribution, comme la moyenne, la variance, la médiane, *etc*.

Plan partie 1

Tests non paramétriques

Principe

Faire un test statistique, sans spécifier la distribution des variables aléatoires observées.

Exemples

- Test d'adéquation de Kolmogorov Smirnov (KS test),
- Test d'adéquation du χ^2 de Pearson,
- Tests de normalité,

Dans tous ces cas, la distribution des variables n'est pas spécifiée sous l'alternative $\rightarrow$ test non paramétrique.

- Test du χ^2 d'indépendance,
- Tests de corrélation (Pearson, Kendall ou Spearman),
- ...

Test d'adéquation de Kolmogorov Smirnov (KS test) I

Description

- Pour un échantillon de v.a. réelles $X_1, \ldots, X_n$ et une fdr F_0 fixée, on veut tester H_0 : "$F = F_0$" contre H_1 : "$F \neq F_0$". On utilise la statistique

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|.$$

- Pour deux échantillons de v.a. réelles $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ on peut aussi tester $H_0 : F_X = F_Y$ contre $H_1 : F_X \neq F_Y$, via la statistique

$$D_{n,m} = \sup_{t \in \mathbb{R}} |\hat{F}_{n,X}(t) - \hat{F}_{m,Y}(t)|.$$

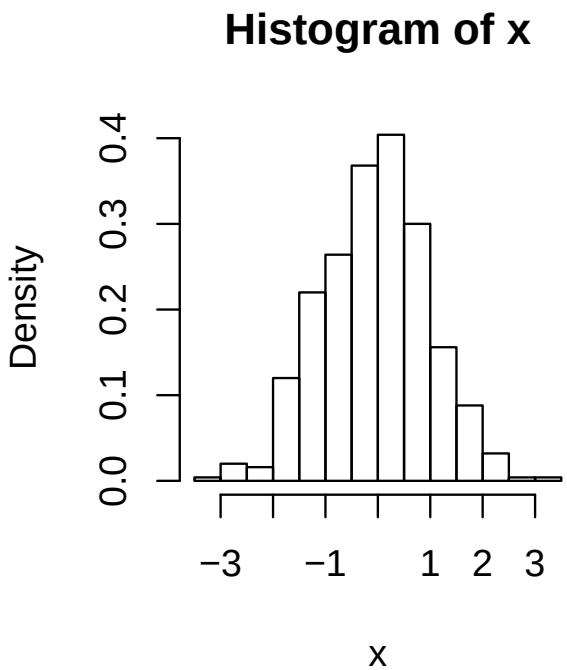
Test d'adéquation de Kolmogorov Smirnov (KS test) II

Propriétés

- Test asymptotique, fondé sur le fait que la distribution limite de $\sqrt{n}D_n$ (resp. $\sqrt{nm/(n+m)}D_{n,m}$) ne dépend pas de la distribution de l'échantillon initial ("asymptotiquement libre en loi").
- Cette distribution est tabulée.
- Test restreint au cas où l'on suppose que la fdr de l'échantillon est continue : variables diffuses.
- Test sensible à des différences à la fois dans la forme et dans la localisation des distributions.

Illustration KS test - Détection d'une différence de localisation I

```
> x <- rnorm(500,0,1)
> y <- rnorm(700,0.15,1)
> par(mfrow=c(1,2))
> hist(x,breaks=10,proba=T)
> hist(y,breaks=10,proba=T)
```



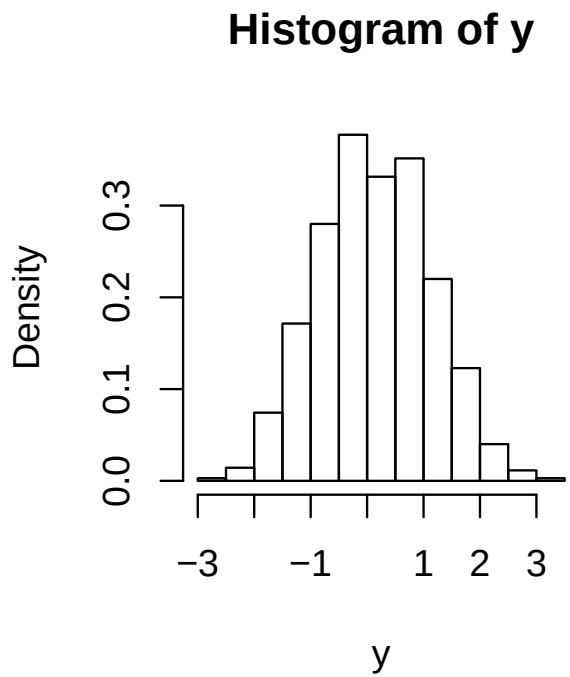


Illustration KS test - Détection d'une différence de localisation II

```
> ks.test(x,y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.0917, p-value = 0.01479
alternative hypothesis: two-sided
```

Illustration KS test - Détection d'une différence de forme I

```
> x <- rnorm(500,0,1)
> y <- rnorm(700,0,1.2)
> par(mfrow=c(1,2))
> hist(x,breaks=10,proba=T)
> hist(y,breaks=10,proba=T)
```

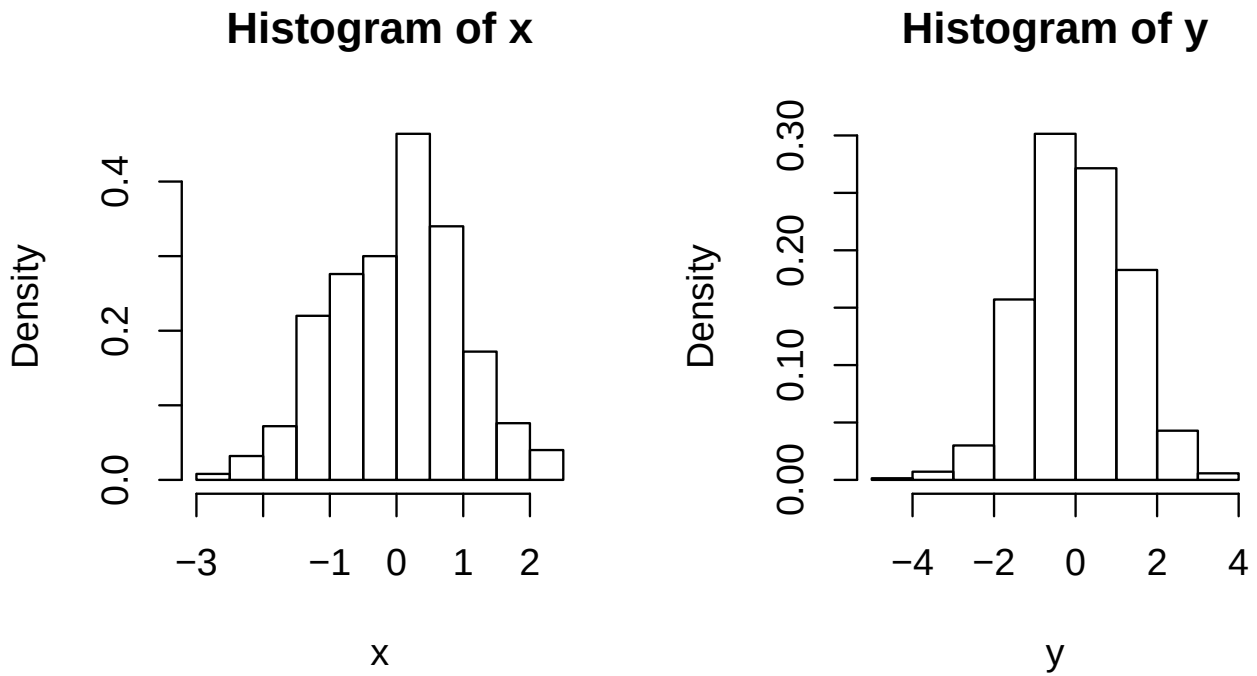


Illustration KS test - Détection d'une différence de forme II

```
> ks.test(x,y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.1063, p-value = 0.002749
alternative hypothesis: two-sided
```

Plan partie 1

Estimation de densité I

Problème

On observe $X_1, \ldots, X_n$ v.a. i.i.d., diffuses, de densité f. On veut estimer f.

Première remarques

- Impossible de faire du max. de vraisemblance sur toutes les densités possibles
$$\sup_f \sum_{i=1}^n \log f(X_i) = +\infty$$
- Nécessité de restreindre la classe de densités considérées.

Quel type d'hypothèses ?

- Régularité : *cf.* Partie 3 de ce cours,
- Monotonie (estimateur de Grenander), Convexité, unimodalité : *cf.* Partie 4 de ce cours.

Estimation de densité II

Qualité de l'estimation

Deux types d'erreur :

- Biais = induit par le choix du modèle. Il traduit la distance de la vraie densité au modèle. Cette erreur diminue lorsqu'on passe d'un modèle paramétrique à un modèle non paramétrique.
- Variance = induite par l'approximation dans le modèle, qui est un espace plus ou moins grand. Cette erreur augmente lorsqu'on passe d'un modèle paramétrique à un modèle non paramétrique.

Nécessité d'un compromis biais/variance.

Plan partie 1

Régression (non paramétrique)

Problème

- On observe une suite de couples $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ avec

$$Y_i = r(X_i) + \epsilon_i$$

où r est une fonction quelconque (régulière) que l'on cherche à estimer.

- Pas abordé dans cet atelier.
- On peut mettre en œuvre les mêmes techniques d'estimation non paramétriques que pour la densité. La différence majeure réside dans la classe de fonctions r qui n'est pas contrainte à avoir une intégrale finie.

Plan partie 1

Estimation en grande dimension

Problème

- Nombre de paramètres $p > n$ nombre d'observations.
- Approches possibles (non exclusives)
  - Modèles creux (ou sparses en anglais) : on fait l'hypothèse qu'un grand nombre de paramètres est en fait nul, mais on ne sait pas lesquels.
  - Approches de type sélection de variables : là encore, on fait l'hypothèse qu'un grand nombre de covariables sont non pertinentes et on cherche à sélectionner celles qui le sont.

Plan partie 1

Bootstrap I

Principe

- Ré-échantilloner les observations pour obtenir des informations sur la population globale dont l'échantillon initial est issu.
- Échantillon bootstrap : construit à partir d'un tirage avec remise de l'échantillon initial, de même taille que l'échantillon initial.
- À partir de B éch. bootstrap, on peut par ex. estimer le biais et la variance d'une statistique S. Ou encore la longueur ou la forme d'intervalles de confiance.

Exemple

Échantillon initial $x_1, \ldots, x_n$, moyenne empirique $\bar{x}$. On tire B échantillons bootstrap $\{x^{\star b} = (x_1^{\star b}, \ldots, x_n^{\star b}), 1 \leq b \leq B\}$. Chaque échantillon bootstrap $x^{\star b}$ a une moyenne empirique $\bar{x}^{\star b}$. Alors,

Bootstrap II

- Si on trace un histogramme des $\{\bar{x}^{\star b}, 1 \leq b \leq B\}$, on approche la distribution de l'estimateur $\bar{x}$ (quand $B \to +\infty$).
- La différence

$$\text{biais}_{\text{boot}} = \bar{x}^{\star} - \bar{x} = \frac{1}{B} \sum_{b=1}^{B} \bar{x}^{\star b} - \bar{x}$$

  approche le biais de l'estimateur $\bar{x}$,
- et l'erreur empirique

$$\text{se}_{\text{boot}} = \left[\sum_{b=1}^{B} \frac{(\bar{x}^{\star b} - \bar{x}^{\star})^2}{B-1} \right]^{1/2}$$

  approche la déviation standard de l'estimateur $\bar{x}$.

Jackknife

Voir la section sur les fonctions d'influence.

Plan partie 1

Rappels sur les fonctions de répartition I

- On note $\mathcal{F}$ l'ensemble des fonctions de répartitions (fdr)

$$\mathcal{F} = \{F : \mathbb{R} \to [0,1]; F \text{ croissante, càdlàg, } \lim_{t\to-\infty} F(t) = 0, \lim_{t\to+\infty} F(t) = 1\}.$$

- Si $F \in \mathcal{F}$, on note dF l'unique mesure de proba associée et on peut définir ainsi la notation $\int h(x)dF(x)$ pour toute fonction $h : \mathbb{R} \to \mathbb{R}$.
- Exemple :
  - si F continue (*i.e.* si dF est une mesure absolument continue) alors en notant $f = F'$ la densité on obtient $\int h(x)dF(x) = \int h(x)f(x)dx$
  - si F constante par morceaux (*i.e.* si dF est une mesure discrète) alors $\int h(x)dF(x) = \sum_{a\in\mathcal{A}} h(a)w_a$ où $\mathcal{A}$ est le support de la mesure et $\{w_a\}_{a\in\mathcal{A}}$ l'ensemble des poids associés.

Rappels sur les fonctions de répartition II

FDR empirique $\hat{F}_n$

Soient $X_1, \ldots, X_n$ v.a.i.i.d. réelles

- $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq t}, \quad \forall t \in \mathbb{R}.$
- Constante par morceaux (croissante, càd-làg)
- Associée à la mesure empirique

$$\mathbb{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot)$$

où δ_x est la masse de Dirac au point x, *i.e.* $\mathbb{P}_n$ est une mesure discrète qui associe le poids $1/n$ à chacune des observations X_i.

- Pour toute fonction h, on a $\int h(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$.

Plan partie 1

Fonctionnelles de la distribution

Une fonctionnelle est une application $T : \mathcal{F} \to \mathbb{R}$.

Exemples

- Moyenne : $F \to \mu(F) = \int x dF(x)$,
- Variance : $F \to \sigma^2(F) = \int (x - \mu(F))^2 dF(x) = \int x^2 dF(x) - (\int x dF(x))^2$,
- Médiane : $F \to m(F) = F^{-1}(1/2)$ et Quantiles : $F \to q_\alpha(F) = F^{-1}(\alpha)$,
- Skewness (ou coefficient d'asymétrie) : $F \to \{\int (x - \mu(F))^3 dF(x)\}/\sigma(F)^{3/2}$.
- $\mathbb{E}(|X_1 - X_2|), \mathbb{P}((X_1, X_2) \in S), \ldots$

Cas particuliers : fonctionnelles linéaires et fonctionnelles de moment

Définitions

- Une fonctionnelle T est dite linéaire s'il existe $a : \mathbb{R} \to \mathbb{R}$ telle que $T : F \to T(F) = \int a(x) dF(x)$.
- Une fonctionnelle T est dite de moment s'il existe un entier $k \geq 1$ et une fonction $\phi : \mathbb{R}^k \to \mathbb{R}$ telle que $T : F \to T(F) = \mathbb{E}(\phi(X_1, \ldots, X_k)) = \int \phi(x_1, \ldots, x_k) dF(x_1) \ldots dF(x_k)$.

Exemples

- Les fonctionnelles linéaires sont des fonctionnelles de moment.
- Moyenne : linéaire et de moment
- $\mathbb{E}(|X_1 - X_2|), \mathbb{P}((X_1, X_2) \in S)$: fonctionnelles de moment.
- Variance, médiane, quantiles, skewness : NON

Estimateurs par substitution I

Principe des estimateurs par substitution (ou "plug-in")

Si $T : F \to T(F)$ est une fonctionnelle alors un estimateur naturel de $T(F)$ est obtenu en substituant l'estimateur $\hat{F}_n$ de F dans l'expression de T, *i.e.* $\hat{T}_n = T(\hat{F}_n)$ est un estimateur naturel de $T(F)$.

Exemples

- Moyenne empirique : $\bar{X}_n = \int x d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n F(X_i)$,
- Variance empirique :

$$\hat{\sigma}_n^2 = \int x^2 d\hat{F}_n(x) - \left(\int x d\hat{F}_n(x)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Estimateurs par substitution II

- Variance empirique (suite) : Estimateur biaisé auquel on peut préférer

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

qui est sans biais.

- Médiane empirique $\hat{m} = \hat{F}_n^{-1}(1/2)$,
- Quantile empirique $\hat{q}_\alpha = \hat{F}_n^{-1}(\alpha)$.
- Statistiques d'adéquation
  - Kolmogorov : $\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$,
  - Cramer von Mises : $\int (\hat{F}_n(x) - F_0(x))^2 dF_0(x)$,
  - Pearson : $\sum_{j=1}^{r} \frac{(\hat{p}_j - p_j^0)^2}{p_j^0}$ où $p_j^0 = F_0((a_j; a_{j+1}])$ et $\hat{p}_j = \hat{F}_n((a_j; a_{j+1}])$. Correspond à $T(F) = \sum_{j=1}^{r} \frac{(F(a_{j+1}) - F(a_j) - p_j^0)}{p_j^0}$.

U et V statistiques I

Estimateurs des fonctionnelles de moment

Soit $T = \mathbb{E}(\phi(X_1, \ldots, X_k))$ une fonctionnelle de moment.
(On peut supposer ϕ symétrique en les coordonnées).

- Son estimateur de substitution est la V-statistique

$$V = T(\hat{F}_n) = \frac{1}{n^k} \sum_{i_1=1}^{n} \ldots \sum_{i_k=1}^{n} \phi(X_{i_1}, \ldots, X_{i_k}).$$

  C'est un estimateur biaisé.

- Un estimateur sans biais de T est la U-statistique

$$U = \binom{n}{k}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \ldots \sum \phi(X_{i_1}, \ldots, X_{i_k}).$$

- La U-stat et la V-stat correspondante ont le même comportement asymptotique.

U et V statistiques II

Exemple

- Différence en moyenne de Gini : $U = \binom{n}{2}^{-1} \sum_{i<j} |X_i - X_j|$

Propriétés des U-statistiques

- Estimateurs sans biais.
- $\mathbb{V}\mathrm{ar}(\sqrt{n}\,U) \to k^2\sigma_1^2$ où
  $\sigma_1^2 = \mathbb{C}\mathrm{ov}(\phi(X, X_2, \ldots, X_k); \phi(X, X_2', \ldots, X_k'))$ et
  $X, X_2, \ldots, X_k, X_2', \ldots, X_k'$ i.i.d. de loi F.
- Si $\sigma_1^2 < +\infty$, alors $\sqrt{n}(U - T(F)) \rightsquigarrow^{\mathcal{L}}_{n\to\infty} \mathcal{N}(0, k^2\sigma_1^2)$ et
  $\sqrt{n}(V - T(F)) \rightsquigarrow^{\mathcal{L}}_{n\to\infty} \mathcal{N}(0, k^2\sigma_1^2)$.

Consistance des estimateurs par substitution I

Principe

- On a vu que $\hat{F}_n$ converge (de diverses façons) vers F.
- Si T est assez régulière, les propriétés de $\hat{F}_n$ se transmettent à $\hat{T}_n = T(\hat{F}_n)$.
- Attention : T est définie sur $\mathcal{F}$: notion de régularité à préciser.

Rappel : Lemme de Slutsky

Soit $(X_n)_{n\geq 0}$ suite de v.a. dans $\mathbb{R}^d$ qui converge en loi vers X et $h : \mathbb{R}^d \to \mathbb{R}^s$ continue. Alors $(h(X_n))_{n\geq 0}$ suite de v.a. dans $\mathbb{R}^s$ qui converge en loi vers $h(X)$.

Continuité d'une fonctionnelle

Un fonctionnelle T est continue au point F si

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0 \Rightarrow T(F_n) - T(F) \to 0.$$

Consistance des estimateurs par substitution II

Exemples de fonctionnelles continues

- Fdr en un point $T : F \mapsto F(x_0)$
- Cramer von Mises $T : F \mapsto \int (F - F_0)^2 dF_0$
- Quantiles $T : F \mapsto F^{-1}(\alpha)$
- Contre-exemple : La moyenne n'est pas continue. En général, les fonctionnelles de moment ne sont pas continues.

Convergence de l'estimateur plug-in (Condition suffisante)

Si $T : \mathcal{F} \to \mathbb{R}$ est continue alors $\hat{T}_n = T(\hat{F}_n)$ converge en proba vers $T(F)$.

Consistance des estimateurs par substitution III

Exemples d'estimateurs par substitution consistants

- Moyenne empirique : $T : F \mapsto \int x dF(x)$ n'est pas continue en tout point mais on a quand même $\bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}(X)$,
- Variance empirique : $\hat{\sigma}_n^2 \to^{\mathbb{P}} \mathbb{V}\mathrm{ar}(X)$
- Médiane empirique $\hat{m} = \hat{F}_n^{-1}(1/2) \to^{\mathbb{P}} F^{-1}(1/2)$,
- Quantile empirique $\hat{q}_\alpha = \hat{F}_n^{-1}(\alpha) \to^{\mathbb{P}} F^{-1}(\alpha)$.

Normalité asymptotique des estimateurs par substitution I

Rappel : Méthode Delta

Si $(X_n)_{n\geq 0}$ suite de v.a. dans $\mathbb{R}^d$ telles qu'il existe $\mu \in \mathbb{R}^d$ et $(a_n)_{n\geq 0}$ suite de réelles avec $a_n(X_n - \mu) \rightsquigarrow^{\mathcal{L}}_{n\to\infty} \mathcal{N}_d(0, \Sigma)$ et si $g : \mathbb{R}^d \to \mathbb{R}^s$ est différentiable au voisinage de μ, alors

$$a_n(g(X_n) - g(\mu)) \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}_s(0, \nabla g(\mu)^\intercal . \Sigma . \nabla g(\mu)).$$

Exemple d'application directe : variance empirique

$$\sqrt{n}(\hat{\sigma}_n^2 - m_2) \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, m_4 - m_2^2),$$

où $m_i = \mathbb{E}[(X_1 - \mathbb{E}X_1)^i]$. (*Indication : écrire un TCL sur le vecteur* $(\bar{X}_n, \overline{X_n^2})$).

Dérivabilité d'une fonctionnelle

C'est la notion de fonction d'influence.

Plan partie 1

Fonction d'influence I

Principe

- Équivalent de la fonction de score en statistique paramétrique.
- C'est une dérivée de la fonctionnelle.
- Pour définir une dérivée, il faut définir un taux d'accroissement. Comme une fonctionnelle T a pour argument $F \in \mathcal{F}$, il faut définir un accroissement élémentaire dans $\mathcal{F}$.

Accroissement élémentaire

$\forall x_0 \in \mathbb{R}$, on note δ_{x_0} la masse de Dirac en x_0 et $G_{\delta_{x_0}}$ la f.d.r. associée à δ_{x_0}. Plus précisément, on a $G_{\delta_{x_0}}(t) = 1_{x_0 \leq t}$ pour tout $t \in \mathbb{R}$.

Fonction d'influence II

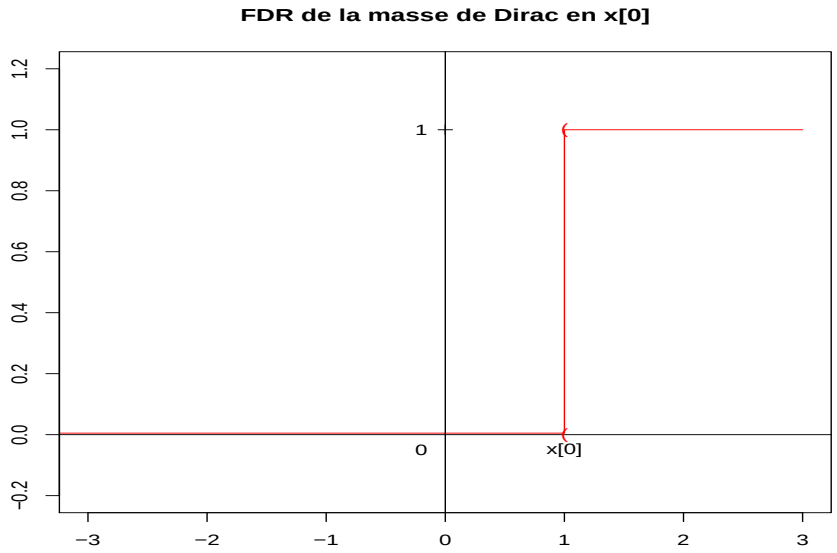


Figure : Fonction de répartition $G_{\delta_{x_0}}$ de la masse de Dirac en x_0.

Fonction d'influence III

Définition

Soit $T : F \to T(F)$ une fonctionnelle. La fonction d'influence de T en F au point x_0 est définie par la limite suivante, si elle existe pour tout $x \in \mathbb{R}$,

$$IF_{T,F}(x_0) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) - T(F)}{\epsilon}.$$

Remarque

Si $F \in \mathcal{F}$ alors pour tout $\epsilon > 0$, on a $(1-\epsilon)F + \epsilon G_{\delta_{x_0}} \in \mathcal{F}$. En effet, c'est une fonction croissante, càdlàg, qui tend vers 0 en $-\infty$ et vers 1 en $+\infty$.

Heuristique

- $IF_{T,F}(x_0)$ mesure la variation limite subie par la fonctionnelle T en la distribution F lors d'une perturbation infinitésimale.
- Notion de contamination, liée à la robustesse statistique.

Fonction d'influence IV

Exemple (fonctionnelle linéaire)

Moyenne $\mu : F \to \mu(F) = \int x dF(x)$.
$\forall \epsilon > 0, \forall x_0 \in \mathbb{R}$, on a

$$\mu((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) = (1-\epsilon)\mu(F) + \epsilon\mu(G_{\delta_{x_0}})$$

car μ est linéaire ! De plus, $\mu(G_{\delta_{x_0}}) = x_0$. Donc on obtient

$$\frac{\mu((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) - \mu(F)}{\epsilon} = \frac{(1-\epsilon)\mu(F) + \epsilon x_0 - \mu(F)}{\epsilon} = x_0 - \mu(F).$$

Ainsi, $IF_{\mu,F}(x_0) = x_0 - \mu(F)$.

Fonction d'influence empirique

Définition

Soit $T : F \to T(F)$ une fonctionnelle. La fonction d'influence empirique de T en F au point x_0 est

$$\hat{IF}_n(x_0) = IF_{T,\hat{F}_n}(x_0).$$

Exemple (suite)

La fonction d'influence empirique associée à la moyenne μ en F au point x_0 est $\hat{IF}_n(x_0) = x_0 - \bar{X}_n$.

Heuristique

La quantité $\hat{IF}_n(X_i)$ mesure la contribution de l'observation X_i à la variation de la statistique $\hat{T}_n$.

Calcul de fonctions d'influence

Proposition

1. *Si* $T, S : \mathcal{F} \to \mathbb{R}$ *sont deux fonctionnelles de fonctions d'influence respectives* $IF_{T,F}$ *et* $IF_{S,F}$ *au point* $F \in \mathcal{F}$ *et si* $\lambda \in \mathbb{R}$ *alors* $\lambda T + S$ *est une fonctionnelle de fonction d'influence* $\lambda IF_{T,F} + IF_{S,F}$ *au point* F *et* $T \times S$ *est une fonctionnelle de fonction d'influence* $T(F) \times IF_{S,F} + S(F) \times IF_{T,F}$ *au point* F.
2. *Si* $\psi : \mathbb{R} \to \mathbb{R}$ *est une fonction dérivable et si* T *est une fonctionnelle de fonction d'influence* $IF_{T,F}$ *au point* F *alors la fonctionnelle* $S = \psi \circ T$ *a pour fonction d'influence au point* F *la fonction* $IF_{S,F} = \psi' \circ T \times IF_{T,F}$.

Application : calcul de la fonction d'influence de la variance

- On a $\sigma^2(F) = \int (x - \mu(F))^2 dF(x) = \int x^2 dF(x) - \mu(F)^2$.
- D'après la prop. précédente $IF_{\sigma^2,F} = IF_{T,F} - 2\mu(F) IF_{\mu,F}$ où $T : F \to T(F) = \int x^2 dF(x)$.
- Or $IF_{\mu,F}(x) = x - \mu(F)$ et T est aussi une fonctionnelle linéaire donc $IF_{T,F}(x) = x^2 - T(F) = x^2 - \int u^2 dF(u)$.
- Donc on obtient

$$\begin{aligned} IF_{\sigma^2,F}(x) &= x^2 - \int u^2 dF(u) - 2\mu(F)(x - \mu(F)) \\ &= (x - \mu(F))^2 - \sigma^2(F). \end{aligned}$$

- NB : On retrouve la forme $IF_{\sigma^2,F}(x) = a_F(x) - \sigma^2(F)$ avec $\sigma^2(F) = \int a_F(x) dF(x)$ et $a_F(x) = (x - \mu(F))^2$, pourtant, σ^2 n'est pas une fonctionnelle linéaire car la fonction a dépend de F.

Construction d'intervalles de confiance pour $T(F)$ I

Exemple de la moyenne $\mu(F) = \int x dF(x)$

- D'après le TCL

$$\sqrt{n}\frac{(\mu(F) - \bar{X}_n)}{\sigma(F)} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1),$$

- On remarque que comme $IF_{\mu,F}(X) = X - \mu(F)$, on a $\sigma^2(F) = \mathbb{V}\mathrm{ar}(X) = \mathbb{V}\mathrm{ar}(IF_{\mu,F}(X))$.
- Mais $\sigma^2(F)$ est inconnue. On l'estime par $\hat{\sigma}_n^2 = \sigma^2(\hat{F}_n)$, ce qui revient à estimer $\sigma^2(F)$ par $\mathbb{V}\mathrm{ar}(\hat{IF}_n(X))$.
- Au final, le Lemme de Slutsky combiné au TCL donne

$$\sqrt{n}\frac{(\mu(F) - \bar{X}_n)}{\sqrt{\mathbb{V}\mathrm{ar}(\hat{IF}_n)}} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1).$$

Construction d'intervalles de confiance pour $T(F)$ II

Théorème (Cas des fonctionnelles linéaires).

Si $T : F \to T(F)$ est une fonctionnelle linéaire, *i.e.* de la forme $T(F) = \int a(x)dF(x)$, alors

i) $IF_{T,F}(x_0) = a(x_0) - T(F)$ et
$\hat{IF}_n(x_0) = a(x_0) - T(\hat{F}_n) = a(x_0) - \frac{1}{n}\sum_{i=1}^n a(X_i)$,

ii) $\forall H \in \mathcal{F}$, on a $T(H) = T(F) + \int IF_{T,F}(x)dH(x)$,

iii) $\mathbb{E}(IF_{T,F}(X)) = \int IF_{T,F}(x)dF(x) = 0$,

iv) Soit
$\tau^2 = \int IF^2_{T,F}(x)dF(x) = \mathbb{E}(IF_{T,F}(X)^2) = \mathbb{V}\mathrm{ar}(IF_{T,F}(X))$
alors on a

$$\tau^2 = \int (a(x) - T(F))^2 dF(x) = \int a^2(x)dF(x) - T(F)^2.$$

De plus, si $\tau^2 < +\infty$, alors
$\sqrt{n}(T(F) - T(\hat{F}_n)) \rightsquigarrow^{\mathcal{L}}_{n\to\infty} \mathcal{N}(0, \tau^2)$.

Construction d'intervalles de confiance pour $T(F)$ III

v) On définit $\hat{\tau}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{IF}_n^2(X_i) = \frac{1}{n}\sum_{i=1}^n [a(X_i) - T(\hat{F}_n)]^2$ estimateur de τ^2, alors on a la convergence

$$\hat{\tau}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \tau^2,$$

et par conséquent

$$\sqrt{n}\frac{(T(F) - T(\hat{F}_n))}{\hat{\tau}_n} \underset{n\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0,1).$$

Construction d'intervalles de confiance pour $T(F)$ IV

Autres cas

- Pour les fonctionnelles de moment, on a vu la normalité asymptotique des V-statistiques.
- En général, il faut démontrer la normalité asymptotique à la main.

Conclusion

- Souvent, la variance asymptotique de l'estimateur $T(\hat{F}_n)$ vaut l'espérance du carré de la fonction d'influence.

Liens avec les statistiques robustes [Hampel 74] I

Statistiques résumées de la fonction d'influence

Estimateur de la forme $\hat{T}_n = T(\hat{F}_n)$.

- Variance asymptotique : $\mathbb{E}(IF_{T,F}(X)^2)$,
- Gross-error sensitivity : $\gamma^\star = \sup_{x \in \mathbb{R}} |IF_{T,F}(x)|$.
  - Mesure la pire influence approchée qu'un niveau de contamination fixé peut avoir sur la valeur de l'estimateur. On peut le voir comme une borne approchée du biais de l'estimateur.
  - Si $\gamma^\star$ est borné, alors la fonctionnelle T est robuste aux valeurs aberrantes. Ex : la médiane, mais pas la moyenne.
  - Rem : Les méthodes de robustification d'un estimateur cherchent souvent à mettre une borne sur $\gamma^\star$. Le prix à payer est une augmentation de la variance limite.

Liens avec les statistiques robustes [Hampel 74] II

- Local shift sensitivity

$$\lambda^\star = \sup_{x \neq y} \frac{|IF_{T,F}(x) - IF_{T,F}(y)|}{|x - y|}$$

Mesure par exemple les effets locaux de l'arrondi ou du regroupement de valeurs sur la fonctionnelle T.

Liens avec estimateur Jackknife I

Principe du Jackknife

- On observe un n-échantillon. Estimateur initial $\hat{\theta}_n^0$ de θ.
- Pour $1 \leq i \leq n$, on construit $\hat{\theta}_{n-1}^{(i)}$ le même estimateur de θ sur les observations privées de la $i$ème.
- On forme les pseudo-valeurs $\hat{\theta}^{\star,i} = n\hat{\theta}_n^0 - (n-1)\hat{\theta}_{n-1}^{(i)}$.
- Estimateur Jackknife $\hat{\theta} = n^{-1} \sum_{i=1}^{n} \hat{\theta}^{\star,i}$ a un biais réduit.

Liens avec estimateur Jackknife II

Jackknife et fonction d'influence empirique

En prenant $\epsilon = -1/(n-1)$ on a

$$\begin{aligned} \hat{IF}_n(X_i) = IF_{T,\hat{F}_n}(X_i) &\simeq \frac{T((1-\epsilon)\hat{F}_n + \epsilon G_{\delta_{X_i}}) - T(\hat{F}_n)}{\epsilon} \\ &= (n-1)[T(\hat{F}_n) - T(\hat{F}^{(i)}_{n-1})] = \hat{T}^{\star,i} - T(\hat{F}_n). \end{aligned}$$

Le Jackknife peut-être vu comme une version à taille d'échantillon finie d'une fonction d'influence empirique
[Miller & Ruppert 64, Huber 72].

Illustration : Fonctions d'influence pour la reconstruction robuste de phylogénies [Bar-Hen *et al.* 08] I

Problématique

- Reconstruction d'arbre phylogéniques à partir de séquences alignées.
- Détection de sites "outliers" qui ont une valeur de la fonction d'influence très négative : ils modifient beaucoup la topologie inférée lorsqu'on les retire de l'analyse.

Notations

- On observe un tableau $\mathbf{X} = (X_{pq})_{1 \leq p \leq s, 1 \leq q \leq n}$ composé de l'alignement de s séquences sur n sites. X_{pq} est le nucléotide au site q de la séquence p.
- Les colonnes de l'alignement sont $\mathbf{X}_h = (X_{1h}, \ldots, X_{sh})$.

Illustration : Fonctions d'influence pour la reconstruction robuste de phylogénies [Bar-Hen *et al.* 08] II

- On suppose un modèle d'évolution des séquences à sites indépendants.
- On s'intéresse à la fonctionnelle log-vraisemblance

$$\hat{T}_n = \sum_{h=1}^{n} \log f(\mathbf{X}_h | \mathcal{T}, \theta_{\mathcal{T}})$$

où $\mathcal{T}$ est l'arbre et $\theta_{\mathcal{T}}$ les paramètres d'évolution le long de l'arbre.

- L'effet d'un site h est mesuré par la quantité

$$\hat{IF}_n(\mathbf{X}_h) = (n-1)[\hat{T}_n - \hat{T}^{(h)}_{n-1}].$$

Illustration : Fonctions d'influence pour la reconstruction robuste de phylogénies [Bar-Hen *et al.* 08] III

Hypothèses biologiques

- Seul un petit nombre de sites modifie significativement l'arbre inféré.
- On s'attend à avoir un grand nombre de sites avec une faible influence positive et un petit nombre avec une forte influence négative. On cherche à détecter ces derniers.

Données

- Alignement de la petite sous-unité du gène rRNA ($n = 1026$) pour $s = 157$ espèces de champignon.

Illustration : Fonctions d'influence pour la reconstruction robuste de phylogénies [Bar-Hen *et al.* 08] IV

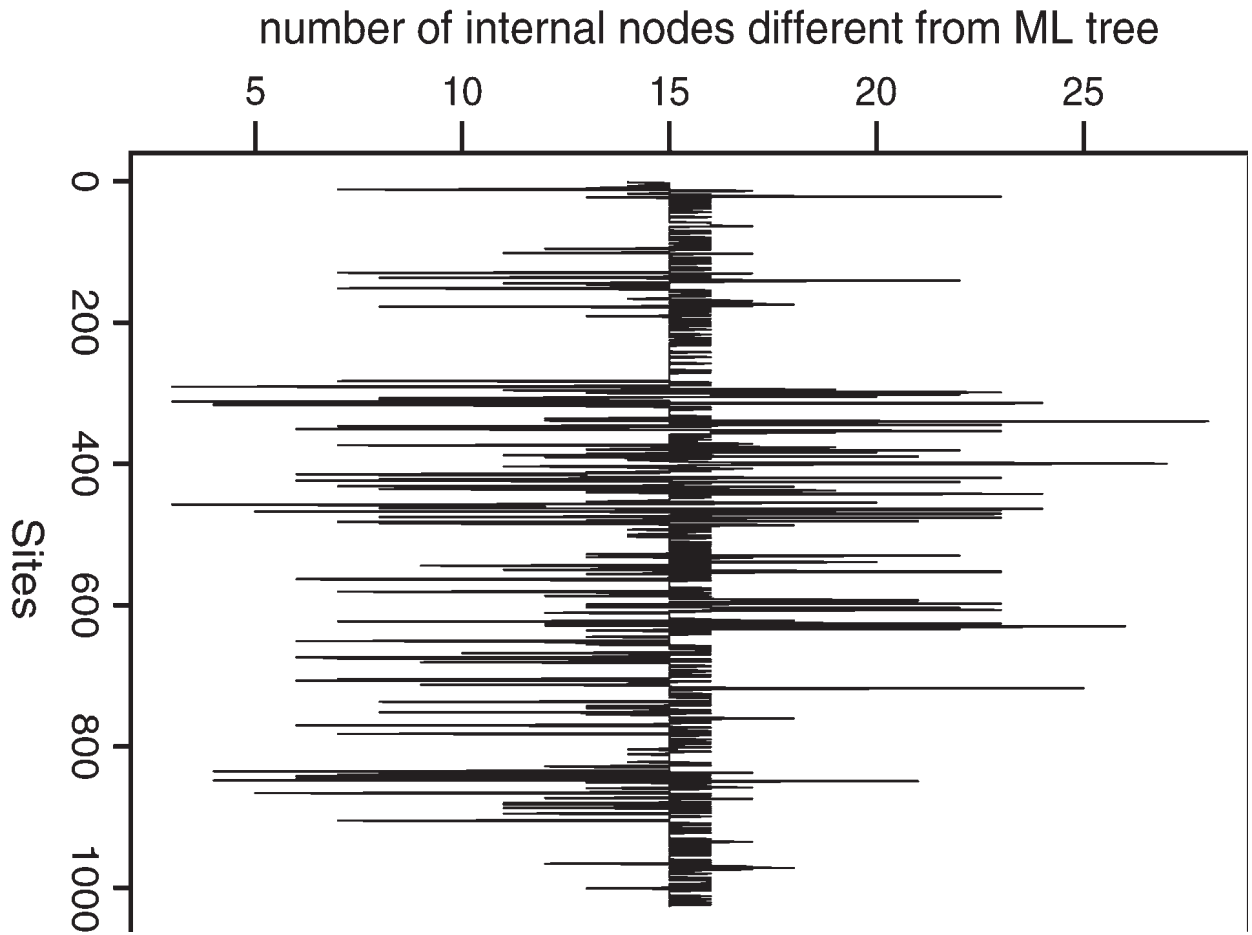


Figure : Nombre de noeuds internes différents de l'arbre initial en fonction du site retiré.

Illustration : Fonctions d'influence pour la reconstruction robuste de phylogénies [Bar-Hen *et al.* 08] V

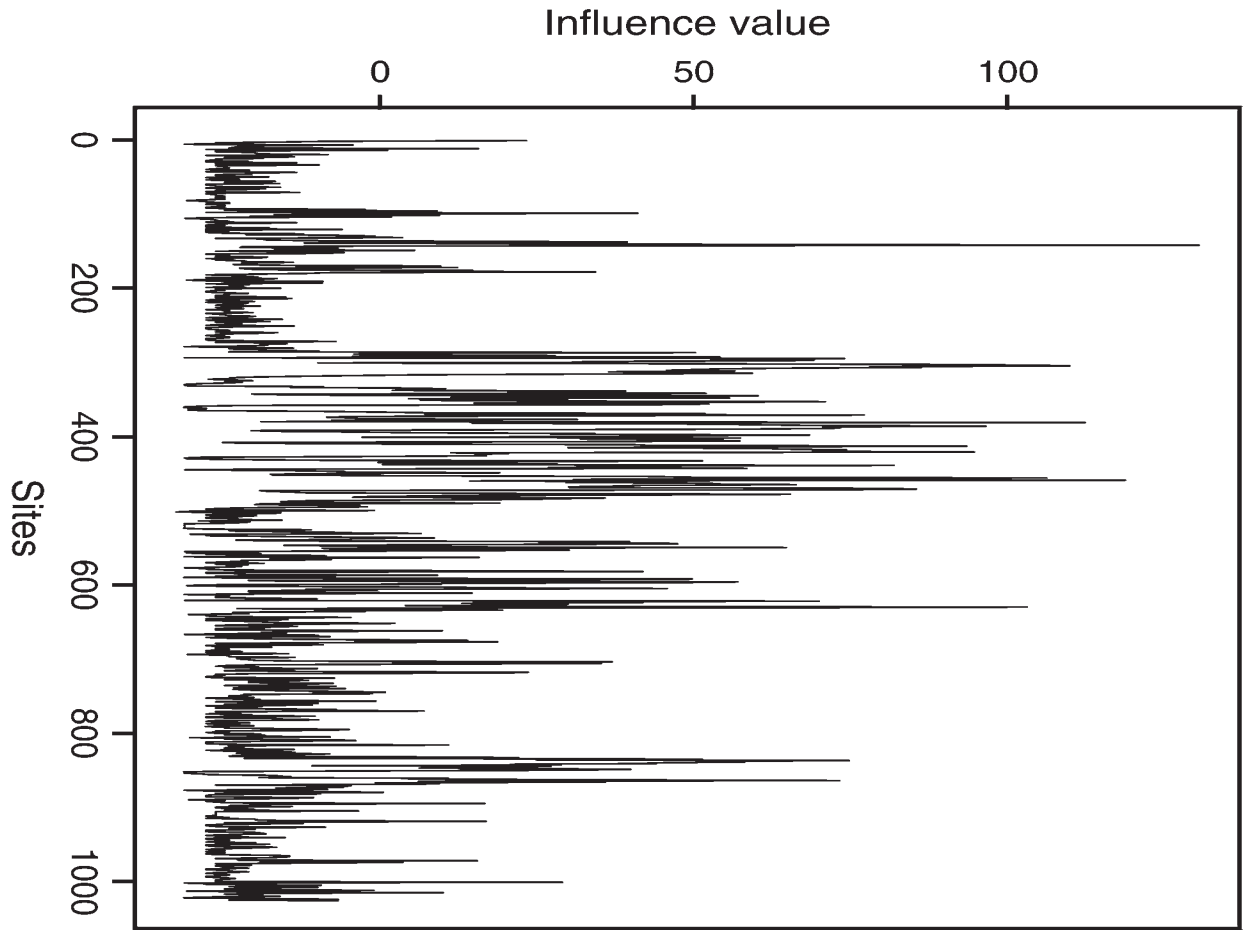


Figure : Valeurs d'influence en fonction du site retiré.

Plan partie 1

Limites des estimateurs par substitution et autres approches

Limites du plug-in

- L'estimateur par substitution $T(\hat{F}_n)$ n'est pas nécessairement "optimal".
- En particulier,
  - $\hat{F}_n$ vise à équilibrer biais et variance pour l'estimation de F,
  - Mais il se peut que $T(\hat{F}_n)$ ne réalise pas l'équilibre biais-variance pour $T(F)$.

TMLE : targeted minimal loss (ou maximum likelihood) estimation

- Stratégie itérative qui permet de construire un estimateur de la fonctionnelle $T(F)$ à partir d'un estimateur initial F_n^0 de F,
- Ici, on cible explicitement l'estimation de $T(F)$ et pas de F !

TMLE : principe

Procédure

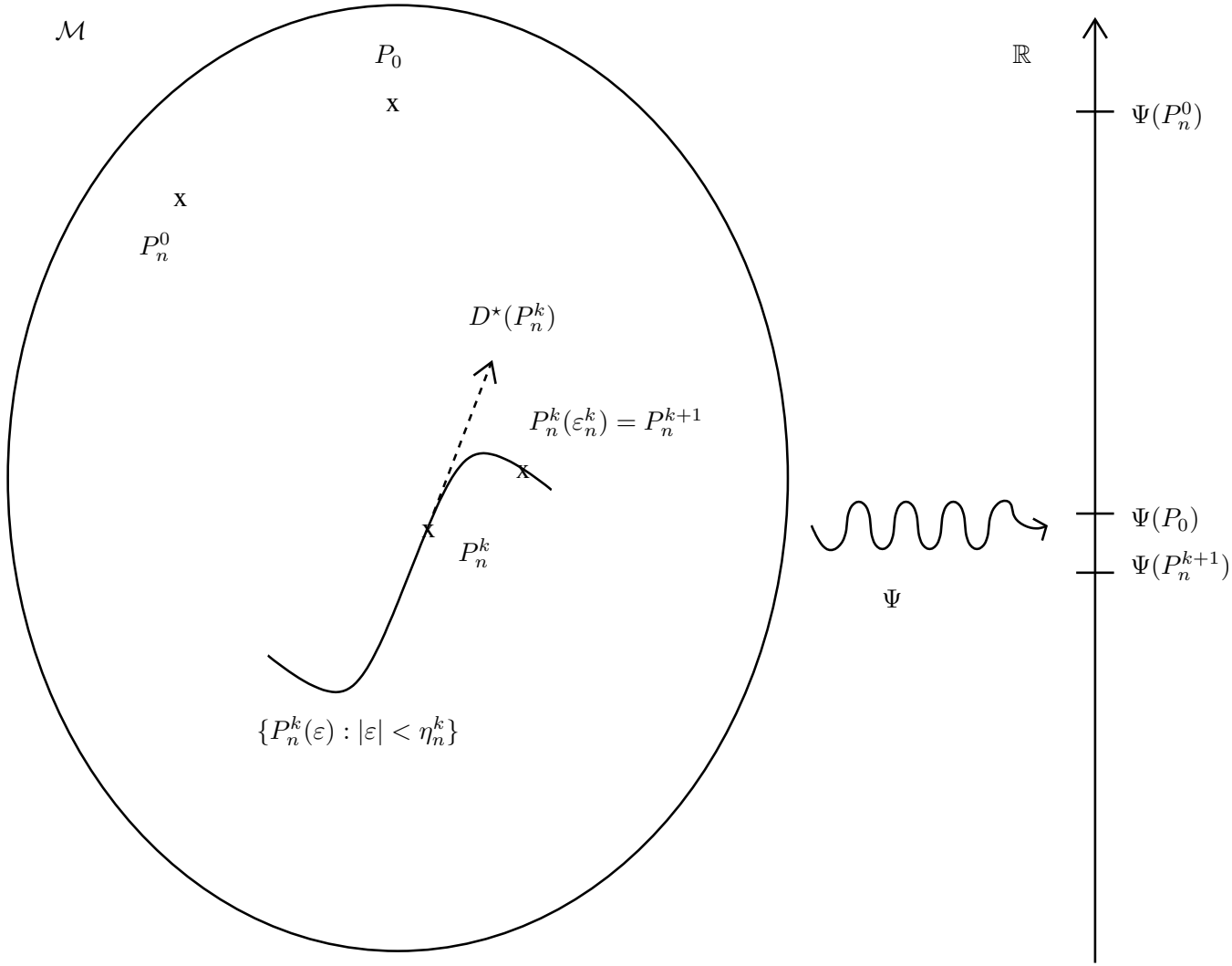
- On part d'un sous-modèle unidimensionel $F_n^0(\epsilon)$ paramétré par $\epsilon \in \mathbb{R}$ dont le score (dérivée) est la fonction d'influence de T en F_n^0,
- On estime le paramètre ϵ par max de vraisemblance ϵ_n^0 (ou minimisation du critère)
- On itère la procédure avec $F_n^1 = F_n^0(\epsilon_n^0)$.

Remarques

- Souvent, une ou deux itérations sont suffisantes pour faire converger la procédure,
- On observe une réduction du biais par la méthode TMLE,
- Il existe des résultats de consistance et de normalité asymptotique pour les estimateurs TMLE.

Pour en savoir plus [Chambaz 11, van der Laan & Rose 11].

TMLE Illustration d'A. Chambaz



Références I


[Bar-Hen *et al.* 08] A. Bar-Hen, M. Mariadassou, M.-A. Poursat, and P. Vandenkoornhuyse.
Influence function for robust phylogenetic reconstructions.
Molecular Biology and Evolution, 25(5) :869–873, 2008.


[Chambaz 11] A. Chambaz.
Estimation et test de l'ordre de lois, de l'importance de variables et de paramètres causaux ; applications biomédicales.
HDR, Université Paris Descartes. 2011


[Hampel 74] Frank R. Hampel.
The influence curve and its role in robust estimation.
J. Amer. Statist. Assoc., 69 :383–393, 1974.


[Huber 72] P. J. Huber.
The 1972 Wald lecture. Robust statistics : A review.
Ann. Math. Statist., 43 :1041–1067, 1972.

Références II


[Miller & Ruppert 64] Miller, Jr., and G. Rupert.
A trustworthy jackknife.
Ann. Math. Statist, 35 :1594–1605, 1964.


[van der Laan & Rose 11] M.J. van der Laan, and S. Rose.
Targeted Learning. Causal Inference for Observational and Experimental Data.
Series : Springer Series in Statistics, 2011.