Statistique de Kolmogorov-Smirnov généralisée

Vernet Elodie

Encadrant : Nicolas Vayatis

ENS de Cachan

24 juin 2010

Le contexte du stage

- Notations :
  - μ une mesure de probabilité sur $\mathbb{R}^d$
  - $\mathcal{C}$ un ensemble de parties mesurables de $\mathbb{R}^d$,
  - $X_1, X_2, ..., X_n$ des variables aléatoires i.i.d. de loi μ,
  - un paramètre $t > 0$,
  - $\hat{\mu_n}(C) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in C\}}$, estimateur de $\mathbb{P}(X_1 \in C) = \mu(C)$.
- Le problème : étudier

$$K(n, \mu, t, \mathcal{C}) = \mathbb{P}\Big\{ \sup_{C \in \mathcal{C}} |\hat{\mu_n}(C) - \mu(C)| > t \Big\}$$

Plan

Généraliser la statistique de Kolmogorov-Smirnov

- F fonction de répartition de μ sur $\mathbb{R}$
- $\hat{F}_n$ fonction de répartition empirique de μ :

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

Généralisation de Kolmogorov-Smirnov :

$$\Delta_n(F) = \mathbb{P}\Big(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \frac{t}{\sqrt{n}} \Big)$$

$$= K(n, \mu, \frac{t}{\sqrt{n}}, \mathcal{C}_0)$$

où $\mathcal{C}_0 = \{C =]-\infty, c], c \in \mathbb{R}\}$.

Inégalités et théorèmes limites

Fonctions de répartition sur $\mathbb{R}$: (Kolmogorov 1933,Dvoretzky-Wolfowitz-Kiefer (1956) , Massart (1990))

Généralisation sur $\mathbb{R}^d$: Vapnik-Chervonenkis (1981).

Résultats asymptotiques

Sur $\mathbb{R}$:

$$\lim_{n\to+\infty} \mathbb{P}\left(sup_{\{x\in\mathbb{R}\}} |\hat{F}_n(x) - F(x)| > \frac{t}{\sqrt{n}} \right) = 2\sum_{k=1}^{+\infty} (-1)^{k-1} e^{-2k^2t^2}$$

Sur $\mathbb{R}^d$: $\forall t > 0\ K(n,\mu,t,\mathcal{C}) \xrightarrow{\mathbb{P}} 0$ avec une CNS.

Résultats non asymptotiques

Sur $\mathbb{R}$:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > t \right) \leq 2e^{-2nt^2}$$

Sur $\mathbb{R}^d$: $\sup_{\mu}(K(n,\mu,t,\mathcal{C})) \leq 4\left(\frac{2en}{V(\mathcal{C})}\right)^{V(\mathcal{C})} \exp\left(-\frac{nt^2}{8}\right)$

Problématique et objectifs du stage

Limites :

Kolmogorov-Smirnov, cadre restreint :

- dans $\mathbb{R}$,
- $\mathcal{C}$ très particulière ;

Vapnik-Chervonenkis, théorie du pire des cas :

- bornes vraies dans le pire des cas
- $\Rightarrow$ majorant non optimal.

Objectif du stage :

étudier $K(n, \mu, t, \mathcal{C})$ dans un cas précis :

- dans $\mathbb{R}$
- avec $\mu \sim \mathcal{U}([0,1])$
- $\mathcal{C}_1(maxa) = \{[0, a,], 0 \leq a < maxa \leq 1\}$.

Difficultés

Simulation d'évènements rares :

- estimer $p = \mathbb{P}(A^j)$ avec A^j dépendant de $X_1^j, X_2^j, ..., X_n^j$
- estimateur des moments : $\hat{p}_N = \frac{1}{N} \sum_{j=1}^{N} \mathbb{P}(A^j)$
- $\mathbb{P}\left\{ \frac{|p-\hat{p}_N|}{p} \leq 0.01 \right\} \geq 95\%$ et $p = 10^{-10} \Rightarrow$ simuler plus de 10^{14} n-uplets $(X_1^j, X_2^j, ..., X_n^j) \Rightarrow$ impossible !

Résolution :

- échantillonnage préférentiel

Calcul d'un suprémum :

- d'une fonction à valeurs continues
- pas d'énumération possible.

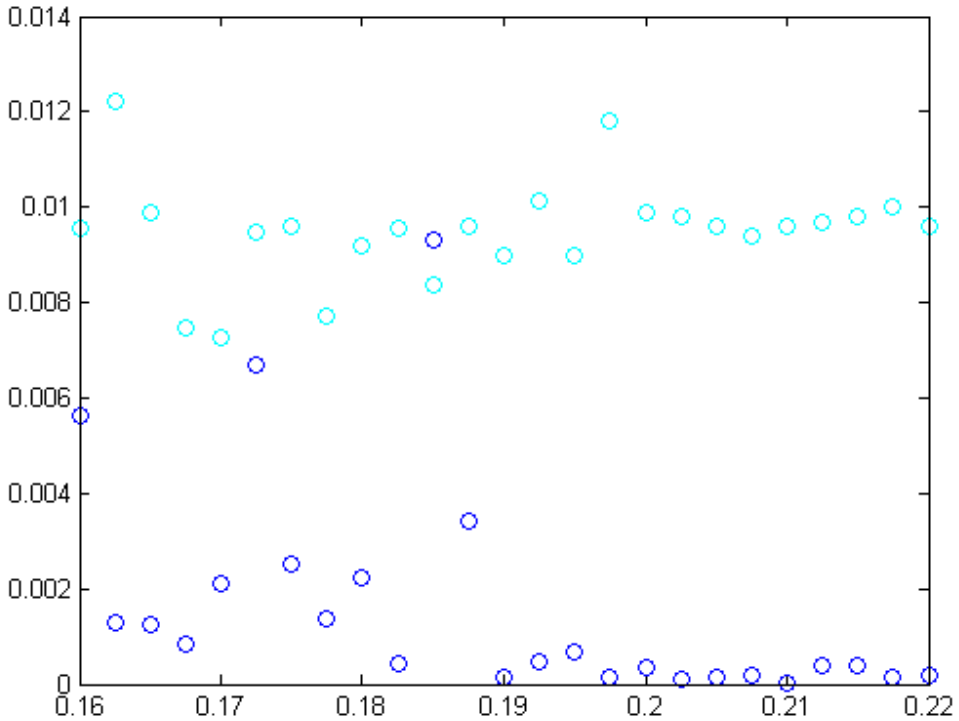
Résolution :

- $\hat{\mu}_n$: nombre fini de valeurs $\Rightarrow$ suprémum $\rightarrow$ maximum.

Expériences numériques

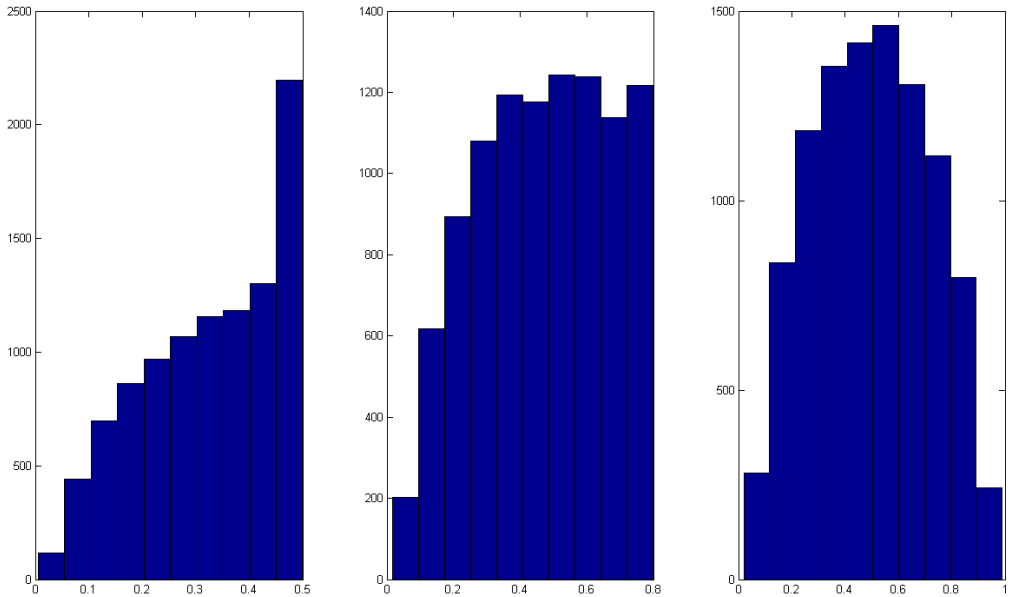
- Expériences réalisées avec matlab,
- nombre de simulations : entre 100 et 10000 (en général 1000) .
- Expériences :
  - variances des estimateurs (comparaison),
  - étude de la répartition des ensembles pour lesquels le suprémum est atteint,
  - comparaison des estimateurs et des majorants théoriques en fonction de n.

Résultats
Échantillonnage préférentiel



$\frac{variance}{moyenne}$ pour des estimateurs de $K(n, \mu, t, \mathcal{C}_0)$ avec et sans échantillonnage préférentiel en fonction de t, $n = 10$.

Résultats
Ensembles critiques



Répartition des a pour lesquels le suprémum est atteint avec $\mathcal{C}_1(maxa) = \{[0, a,], 0 \leq a < maxa \leq 1\}$.
$maxa = 0.5, 0.8, 1$, estimateur : 0.81 ,0.84, 0.84.

Conclusion et perspectives

Conclusion :

- échantillonnage préférentiel utile pour des petites probabilités (≤ 0.02),
- forte dépendance de $K(n, \mu, t, \mathcal{C})$ et des ensembles critiques,
- comportement à décroissance exponentielle.

Perspectives :

- étudier plus en détail la dépendance de $K(n, \mu, t, \mathcal{C})$ et les ensembles critiques,
  - calcul par régression linéaire de la complexité de $\mathcal{C}$ (avec théorie de Vapnik-Chervonenkis);
- étudier un autre exemple sur $\mathbb{R}^2$:
  - $\mu \sim \mathcal{U}([0,1] \times [0,1])$, $\mathcal{C} = \{[0, a] \times [0, b], a, b \in \mathbb{R}\}$ les orthants.

Bibliographie I

James Antonio Bucklew.
Introduction to Rare Event Simulation.
Springer, 2004.

R. Beran et P. W. Millar.
Confidence sets for a multivariate distribution.
The annals of Statistics, 14:431–443, 1986.

Gábor Lugosi.
Concentration of measure inequalities, notes de machine learning summer school 2003, 2005.

C. Read B. Vidakovic et N. L. Johnson S. Kotz, N. Balakrishnan.
Encyclopedia of Statistical Sciences, volume 2.
John Wiley & Sons, 1988.

Nicolas Vayatis.
Thèse : Inégalités de Vapnik-Chervonenkis et mesures de complexité, 2000.