A universal procedure for aggregating estimators
Notes

Lucas BROUX

Janvier 2019

Table des matières

1 Introduction - Théorème 1

Ces notes reprennent l'article [1].

Nous considérons le modèle "bruit blanc gaussien", dans lequel on observe la réalisation d'une fonction inconnue par dessus laquelle s'ajoute un bruit. Formellement, soit

$$\begin{cases} d \in \mathbb{N}, \\ \mathcal{D}_0 := [0,1]^d, \\ W \text{ le processus de Wiener standard dans } \mathbb{R}^d, \\ \epsilon \in (0,1) \text{ correspondant au niveau de bruit.} \end{cases} \tag{1}$$

On observe

$$\mathcal{Y}_\epsilon := \{Y_\epsilon(t), t \in \mathcal{D}_0\}. \tag{2}$$

Où

$$Y_\epsilon(dt) = f(t)\,dt + \epsilon W(dt). \tag{3}$$

On suppose donnés un certain nombre d'estimateurs de f sous la forme $\mathcal{F}_\Theta = \{f_\theta, \theta \in \Theta\}$, paramétrisés par un ensemble Θ.

Idéal. *Trouver :*

$$\underset{f_\theta \in \mathcal{F}_\Theta}{\arg\min} \underbrace{\mathbb{E}_f\left[\|f_\theta - f\|_p\right]}_{=:\mathcal{R}_p[f_\theta; f]}. \tag{4}$$

La difficulté du problème vient du fait que la métrique de risque dépend de la fonction inconnue f. En pratique, on remplace donc cet idéal par le problème suivant :

Objectif. *Trouver un estimateur $f_{\hat{\theta}} \in \mathcal{F}_\Theta$ tel que*

$$\mathcal{R}_p\left[f_{\hat{\theta}}; f\right] \leq C \inf_{\theta \in \Theta} \mathcal{R}_p\left[f_\theta; f\right] + r_\epsilon. \tag{5}$$

Idée 1 (Cours)**.** *On suppose*

- $p = 2$,
- $f_{theta} = LSE(S_\theta)$ *où* S_θ *est un convexe fermé de* $\mathbb{L}_2(\mathcal{D}_0)$.

Dans ce cadre, on utilise

- *La structure euclidienne de* $\mathbb{L}_2(\mathcal{D}_0)$ *pour maitriser le risque.*
- *Une inégalité de concentration gaussienne pour maitriser le processus isonormal.*

Sous hypothèses supplémentaires, on peut montrer une inégalité de type (5).

Idée 2 (Papier)**.** *On ne fait pas d'hypothèses sur les f_θ. Pour simplifier on suppose* $\mathcal{F}_\Theta = \mathcal{F}_{\underbrace{\{1, \cdots, N\}}_{=:I_N}}$ *; et* $p \in [1, +\infty]$.

Explorons désormais ce cadre.

1.1 Le théorème général

Question : comment quantifier la distance de f à f_i ? L'idée est de déléguer la question à un ensemble de fonctions tests. On considère donc une famille de fonctions à ajuster ultérieurement

$$\Psi := \{\psi : \mathcal{D}_0 \to \mathbb{R}\} \tag{6}$$

Et on teste la pertinence de f_i en calculant

$$\begin{aligned}\Delta_i(\psi) &:= \int_{\mathcal{D}_0} \psi(t)\, Y_\epsilon(dt) - \int_{\mathcal{D}_0} \psi(t)\, f_i(t)\, dt \\ &= \int_{\mathcal{D}_0} \psi(t)\,[f(t) - f_i(t)]\, dt + \epsilon \underbrace{\int_{\mathcal{D}_0} \psi(t)\, W(dt)}_{=:\underbrace{Z(\psi)}_{Gaussien}}.\end{aligned} \tag{7}$$

Ainsi, mis à part le terme de bruit $Z(\psi)$, contrôler $\Delta_i(\psi)$ uniformément en ψ, c'est contrôler la "distance" de f à f_i. Autrement dit, sauf en cas de grandes déviations de Z, Δ permet de bien contrôler la "distance" entre f et f_i.

Rigoureusement, on fixe $\delta > 0$ et on considère

$$\chi := \chi(\delta, \Psi) := \min\left\{ x > 0, \mathbb{P}\left[\underbrace{\sup_{\psi\in\Psi} \frac{|Z(\psi)|}{\|\psi\|_2} \geq x}_{=:A_x^c}\right] \leq \delta \right\}. \tag{8}$$

Ainsi sur l'événement A_χ, de probabilité $\geq 1-\delta$, on a

$$\sup_{\psi\in\Psi} \frac{|Z(\psi)|}{\|\psi\|_2} \leq \chi \tag{9}$$

Donc sur A_χ,

$$|\Delta_i(\psi)| \leq \left| \int_{\mathcal{D}_0} \psi(t)\,[f(t) - f_i(t)]\, dt \right| + \epsilon\chi \|\psi\|_2. \tag{10}$$

Et donc sur A_χ,

$$\begin{aligned}|\Delta_i(\psi)| - \epsilon\chi\|\psi\|_2 &\leq \int_{\mathcal{D}_0} |\psi(t)\,[f(t) - f_i(t)]|\, dt \\ &\overbrace{\leq}^{Holder} \|\psi\|_q \|f - f_i\|_p.\end{aligned} \tag{11}$$

Conclusion : sauf sur un événement de probabilité $\leq \delta$,

$$\underbrace{\sup_{\psi\in\Psi}\left(\frac{|\Delta_i(\psi)|-\epsilon\chi\|\psi\|_2}{\|\psi\|_q}\right)}_{=:\hat{M}_i,\text{ connu par observation}} \leq \underbrace{\|f-f_i\|_p}_{\text{inconnu, quantité voulue}}. \tag{12}$$

Cela nous dicte notre procédure :

— On choisit $\hat{i} := \arg\min_{i\in I_N} \hat{M}_i$.

— On retourne $\hat{f} := f_{\hat{i}}$.

Evaluons la qualité de notre procédure. Pour cela, on écrit

$$\left\|f-\hat{f}\right\|_p = \left\|f-\hat{f}\right\|_p \mathbb{1}_{A_\chi} + \left\|f-\hat{f}\right\|_p \mathbb{1}_{A_\chi^c}. \tag{13}$$

Donc

$$\begin{aligned}\mathcal{R}_p\left[\hat{f};f\right] &= \mathbb{E}_f\left[\left\|f-\hat{f}\right\|_p \mathbb{1}_{A_\chi}\right] + \mathbb{E}_f\left[\left\|f-\hat{f}\right\|_p \mathbb{1}_{A_\chi^c}\right] \\ &\leq \mathbb{E}_f\left[\left\|f-\hat{f}\right\|_p \mathbb{1}_{A_\chi}\right] + \max_{i\in I_N}\left(\|f\|_p + \|f_i\|_p\right)\delta.\end{aligned} \tag{14}$$

Ainsi, on veut contrôler $\left\|f-\hat{f}\right\|_p$ en fonction de

$$\min_{i\in I_N}\|f-f_i\|_p =: \|f-f_{i^*}\|_p. \tag{15}$$

Or sous A_χ,

$$\left\|f-\hat{f}\right\|_p \leq \|f-f_{i^*}\|_p + \left\|f_{i^*}-\hat{f}\right\|_p. \tag{16}$$

Reste donc à contrôler $\left\|f_{i^*}-\hat{f}\right\|_p$. C'est le moment de demander des comptes à nos fonctions tests !

Définition 1. *On dit que Ψ est (γ,p)-good si pour tout $i\neq j\in I_N$, il existe $\psi_{i,j}\in\Psi$ tel que*

$$\left|\|f_i-f_j\|_p - \int_{\mathcal{D}_0}\psi_{i,j}(t)\left[f_i(t)-f_j(t)\right]dt\right| \leq \gamma. \tag{17}$$

A partir de maintenant, on suppose que Ψ est (γ, p)-good. Alors sous A_χ,

$$\begin{aligned}
\left\|f_{i^*} - \hat{f}\right\|_p &= \left\|f_{i^*} - f_{\hat{i}}\right\|_p \\
&\leq \left|\int_{\mathcal{D}_0} \psi_{i^*,\hat{i}}(t)\left[f_{i^*}(t) - f_{\hat{i}}(t)\right] dt\right| + \gamma \\
&= \left|\Delta_{i^*}\left(\psi_{i^*,\hat{i}}\right) - \Delta_{\hat{i}}\left(\psi_{i^*,\hat{i}}\right)\right| + \gamma \\
&\leq \left|\Delta_{i^*}\left(\psi_{i^*,\hat{i}}\right)\right| + \left|\Delta_{\hat{i}}\left(\psi_{i^*,\hat{i}}\right)\right| + \gamma \\
&= \left(\frac{\left|\Delta_{i^*}\left(\psi_{i^*,\hat{i}}\right)\right| - \epsilon\chi\left\|\psi_{i^*,\hat{i}}\right\|_2}{\left\|\psi_{i^*,\hat{i}}\right\|_q} + \frac{\left|\Delta_{\hat{i}}\left(\psi_{i^*,\hat{i}}\right)\right| - \epsilon\chi\left\|\psi_{i^*,\hat{i}}\right\|_2}{\left\|\psi_{i^*,\hat{i}}\right\|_q}\right)\left\|\psi_{i^*,\hat{i}}\right\|_q + \gamma + 2\epsilon\chi\left\|\psi_{i^*,\hat{i}}\right\|_2 \\
&\leq \left(\hat{M}_{i^*} + \hat{M}_{\hat{i}}\right)\left\|\psi_{i^*,\hat{i}}\right\|_q + \gamma + 2\epsilon\chi\left\|\psi_{i^*,\hat{i}}\right\|_2 \\
&\leq 2\hat{M}_{i^*}\left\|\psi_{i^*,\hat{i}}\right\|_q + \gamma + 2\epsilon\chi\left\|\psi_{i^*,\hat{i}}\right\|_2 \\
&\leq 2\hat{M}_{i^*}\max_{\psi\in\Psi^*}\left(\|\psi\|_q\right) + \gamma + 2\epsilon\chi\max_{\psi\in\Psi^*}\left(\|\psi\|_2\right) \\
&\leq 2\left\|f - f_{i^*}\right\|_p \max_{\psi\in\Psi^*}\left(\|\psi\|_q\right) + \gamma + 2\epsilon\chi\max_{\psi\in\Psi^*}\left(\|\psi\|_2\right).
\end{aligned} \tag{18}$$

D'où :

Théorème 1. *On fixe*

$$\left|\begin{array}{l}
d \in \mathbb{N}, \\
\mathcal{D}_0 := [0,1]^d, \\
W \text{ le processus de Wiener standard dans } \mathbb{R}^d, \\
\epsilon \in (0,1) \text{ correspondant au niveau de bruit} \\
\delta \in (0,1) \\
\gamma > 0 \\
p \in [1, +\infty].
\end{array}\right. \tag{19}$$

On définit

$$\left|\begin{array}{l}\Theta := \{1, \cdots N\} =: I_N \\ \mathcal{F}_{I_N} = \{f_i, i \in I_N\} \\ \mathcal{G}_{I_N} := \{f_i - f_j, i \neq j\} \\ \Psi_{I_N} := \{\psi_{i,j}, i \neq j\} \\ i^* := \underset{i \in I_N}{\arg\min} \|f - f_i\|_p \\ \Psi^*_{I_N} := \{\psi_{i^*,i}, i \neq i^*\}.\end{array}\right. \tag{20}$$

Notre procédure est de calculer

$$\left|\begin{array}{l}\Delta_i(\psi) := \displaystyle\int_{\mathcal{D}_0} \psi(t)\, Y_\epsilon(dt) - \int_{\mathcal{D}_0} \psi(t)\, f_i(t)\, dt \quad \textit{pour } i \in I_N, \psi \in \Psi \\ \chi := \min\left\{\chi > 0, \mathbb{P}\left[\max_{\psi \in \Psi_{I_N}} \dfrac{|Z(\psi)|}{\|\psi\|_2} \geq \chi\right] \leq \delta\right\} \\ \hat{M}_i := \max_{\psi \in \Psi_{I_N}} \left\{\dfrac{|\Delta_i(\psi)| - \epsilon\chi\|\psi\|_2}{\|\psi\|_q}\right\} \\ \hat{i} := \underset{i \in I_N}{\arg\min}\, \hat{M}_i \\ \hat{f} := f_{\hat{i}}.\end{array}\right. \tag{21}$$

On suppose que Ψ_{I_N} est (γ, p)-good par rapport à $\mathcal{G}_{I_N}$.
Alors :

$$\begin{aligned}\mathcal{R}_p\left[\hat{f}; f\right] \leq &\left(2 \max_{\psi \in \Psi^*_{I_N}} \|\psi\|_q + 1\right) \min_{i \in I_N} \|f - f_i\|_p \\ &+ 2\chi\epsilon \max_{\psi \in \Psi^*_{I_N}} \|\psi\|_2 + \gamma + \left(\|f\|_p + \max_{i \in I_N} \|f_i\|_p\right)\delta.\end{aligned} \tag{22}$$

1.2 Cas particuliers

A partir de ce résultat, plusieurs questions se posent :

— Quel ensemble Ψ conviennent ? Pour quels γ ?

— Peut-on contrôler χ en fonction de δ ?

Dans la suite, on suppose :

— $\delta = \epsilon$.

— $\max\left\{\|f\|_p, \|f_1\|_p, \cdots, \|f_N\|_p\right\} := L < +\infty$.

Commençons par résoudre le conflit entre χ et δ : notons que par définition de (γ, p)-good, on peut choisir $|\Psi| = \frac{N(N-1)}{2}$. Donc

$$\begin{aligned}
\mathbb{P}\left[\sup_{\psi\in\Psi}\frac{|Z(\psi)|}{\|\psi\|_2}\geq x\right] &= \mathbb{P}\left[\bigcup_{\psi\in\Psi}\left\{\frac{|Z(\psi)|}{\|\psi\|_2}\geq x\right\}\right] \\
&\leq \sum_{\psi\in\Psi}\mathbb{P}\left[\underbrace{\frac{|Z(\psi)|}{\|\psi\|_2}}_{\sim\mathcal{N}(0,1)}\geq x\right] \\
&= |\Psi|\, e^{-\frac{x^2}{2}} \\
&\leq N^2 e^{-\frac{x^2}{2}}.
\end{aligned} \tag{23}$$

Ainsi notre ϵ-niveau correspond à

$$\chi = \sqrt{2\log\left(\frac{N^2}{\epsilon}\right)}. \tag{24}$$

En outre, on peut expliciter plusieurs ensembles Ψ intéressants :

Proposition 1. *Soit $p \in [1, +\infty)$, alors*

$$\Psi_{I_N} := \left\{\psi_{i,j} := \frac{|(f_i - f_j)(.)|^{p-1}}{\|f_i - f_j\|_p^{p-1}} sgn\left((f_i - f_j)(.)\right), i \neq j\right\}. \tag{25}$$

est $(0, p)$-good. Et en outre, pour tout $\psi \in \Psi$, $\|\psi\|_q = 1$.

Dès lors, la conclusion du théorème implique

$$\begin{aligned}
\mathcal{R}_p\left[\hat{f}; f\right] \leq & 3\min_{i\in I_N}\|f - f_i\|_p \\
& + 2\epsilon\sqrt{2\log\left(\frac{N^2}{\epsilon}\right)} \underbrace{\max_{\psi\in\Psi^*_{I_N}}\|\psi\|_2}_{} + 0 + 2L\epsilon. \\
& \leq \underbrace{\begin{cases} 1 & \text{si } 1\leq p\leq 2 \\ \max_{i\neq i^*}\left(\frac{\|f_{i^*} - f_i\|_{2p-2}}{\|f_{i^*} - f_i\|_p}\right)^{p-1} & \text{si } 2<p<+\infty \end{cases}}_{=:Q_1(p)}
\end{aligned} \tag{26}$$

Corollaire 1. *On suppose les hypothèses du théorème 1, et en outre* $\max\left\{\|f\|_p, \|f_1\|_p, \cdots, \|f_N\|_p\right\} := L < +\infty$ *et $p \in [1, +\infty)$, alors :*

$$\mathcal{R}_p\left[\hat{f}; f\right] \leq 3\min_{i\in I_N}\|f - f_i\|_p + 2\epsilon\sqrt{2\log\left(\frac{N^2}{\epsilon}\right)}Q_1(p) + 2L\epsilon. \tag{27}$$

1.3 Questions

Quelques questions se posent :

— Peut-on améliorer cela lorsque $p = 2$? Réponse : oui, c'est facile : en exploitant la structure euclidienne sous-jacente.

— Existe-t'il une procédure lorsque ϵ n'est pas supposé connu ? Réponse : oui, c'est facile : en modifiant légèrement notre procédure.

— Peut-on faire de l'agrégation convexe ? Réponse : oui, ce n'est pas trop difficile, on peut expliciter une procédure similaire.

— Le théorème est-il optimal ? Réponse : ce n'est pas évident à priori...

2 Inégalité minimax

Théorème 2. *On se place dans le cas $N > 3, p \in (2, +\infty]$. Alors il existe une famille de fonctions $\bar{\mathcal{F}}_{I_N} := \left\{\bar{f}_i, i \in I_N\right\}$ sur $\mathcal{D}_0$ telle que*

(i) $\max_{i \in I_N} \left\|\bar{f}_i\right\|_p \leq L$.

(ii) pour tous $\epsilon \leq L \left(N \log\left(N\right)\right)^{-\frac{1}{2}}$,

$$\inf_{\tilde{f}: \mathcal{Y}_\epsilon \to \bar{\mathcal{F}}_{I_N}} \max_{f \in \bar{\mathcal{F}}_{I_N}} \left(\mathcal{R}_p\left[\tilde{f}; f\right] - \min_{i \in I_N} \left\|f - \bar{f}_i\right\|_p \right) \geq cK_p \epsilon \sqrt{\log\left(N-1\right)}. \tag{28}$$

Démonstration. L'idée est de considérer $B_i, i = 1, \cdots, N$ des boréliens disjoints de $\mathcal{D}_0$ de mesure de Lebesgue $h \leq \frac{1}{N}$ à ajuster ; puis on pose $\bar{f}_i := L \mathbb{1}_{B_i}$. Alors l'ensemble $\bar{\mathcal{F}}_{I_N} := \left\{\bar{f}_i, i \in I_N\right\}$ convient. En effet, (i) est vraie par construction. En outre, notons

$$s := \left\|\hat{f}_i - \hat{f}_j\right\|_p = (2h)^{\frac{1}{p}} L. \tag{29}$$

Fixons $\tilde{f} : \mathcal{Y}_\epsilon \to \bar{\mathcal{F}}_{I_N}$. L'inégalité de Markov :

$$\begin{aligned} \max_{f \in \bar{\mathcal{F}}_{I_N}} \left(\mathcal{R}_p\left[\tilde{f}; f\right] - \min_{i \in I_N} \left\|f - \bar{f}_i\right\|_p \right) &= \max_{i \in I_N} \left(\mathcal{R}_p\left[\tilde{f}; f\right] \right) \\ &\geq \frac{s}{2} \max_{i \in I_N} \mathbb{P}_i \left[\left\|\tilde{f} - \hat{f}_i\right\|_p \geq \frac{s}{2} \right] \\ &\geq \frac{s}{2} \max_{i \in I_N} \mathbb{P}_i \left[\underbrace{\tilde{i} \neq i}_{=: A_i} \right]. \end{aligned} \tag{30}$$

Nous concluons avec une conséquence du lemme de Birgé :

Lemme 1 (Birgé)**.** *Soit* $(\mathbb{P}_i)_{1\leq i\leq N}$ *une famille de probabilités et* $(A_i)_{1\leq i\leq N}$ *une famille d'événements disjoints. Soit* $K := \frac{1}{N-1}\sum_{i=2}^{N} K(\mathbb{P}_i, \mathbb{P}_1)$ *alors*

$$\max_{i\in I_N} \mathbb{P}_i(A_i) \geq 1 - \kappa \wedge \left(\frac{K}{\log(N)}\right). \tag{31}$$

Où $\kappa = \frac{2e}{2e+1}$ *convient.*

Ici, nous pouvons expliciter les entropies mutuelles grâce à la formule de Cameron-Martin (cf cours p. 29) : pour $i \neq j$:

$$K(\mathbb{P}_i, \mathbb{P}_j) = \frac{\left\|\bar{f}_i - \bar{f}_j\right\|_2^2}{2\epsilon^2} = \frac{hL^2}{\epsilon^2}. \tag{32}$$

Donc

$$\max_{i\in I_N} \mathbb{P}_i\left[\underbrace{\tilde{i} \neq i}_{=:A_i}\right] \geq 1 - \kappa \wedge \left(\frac{hL^2}{\epsilon^2 \log(N)}\right). \tag{33}$$

On choisit $h = \frac{\epsilon^2 \log(N)}{L^2}$ (ce qui est possible dès que $\epsilon \leq L(N\log(N))^{-\frac{1}{2}}$) alors

$$\max_{i\in I_N} \mathbb{P}_i\left[\underbrace{\tilde{i} \neq i}_{=:A_i}\right] \geq 1 - \kappa \geq \frac{1}{7}. \tag{34}$$

Réinjectant, on obtient le résultat voulu.

□

3 Application numérique

3.1 Cadre

On observe

$$Y := \underbrace{\mu}_{\in\mathbb{R}^n} + \epsilon \underbrace{W}_{\sim\mathcal{N}(0,\Sigma)}. \tag{35}$$

La procédure mise en évidence ci-dessus s'adapte à ce cadre.

Pour notre expérience :

— $n = 1000$.

— $\Sigma = Id$.

— On fixe K, et on calcule μ en tirant K composantes du vecteur selon une loi $\mathcal{N}(0, \Sigma)$ et les autres composantes sont nulles.

— On tire deux variables : $Y_1 \sim \mathcal{N}\left(\mu, \epsilon_1^2 \Sigma\right)$ et $Y_2 \sim \mathcal{N}\left(\mu, \epsilon_2^2 \Sigma\right)$.

— Nos estimateurs sont 10 estimateurs de projection et 10 estimateurs de seuil.

3.2 Résultats

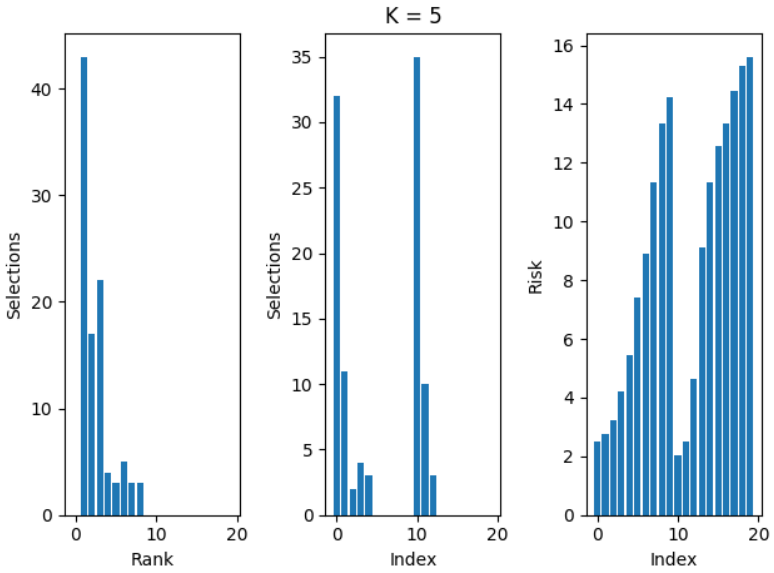

FIGURE 1 – $K = 5$.

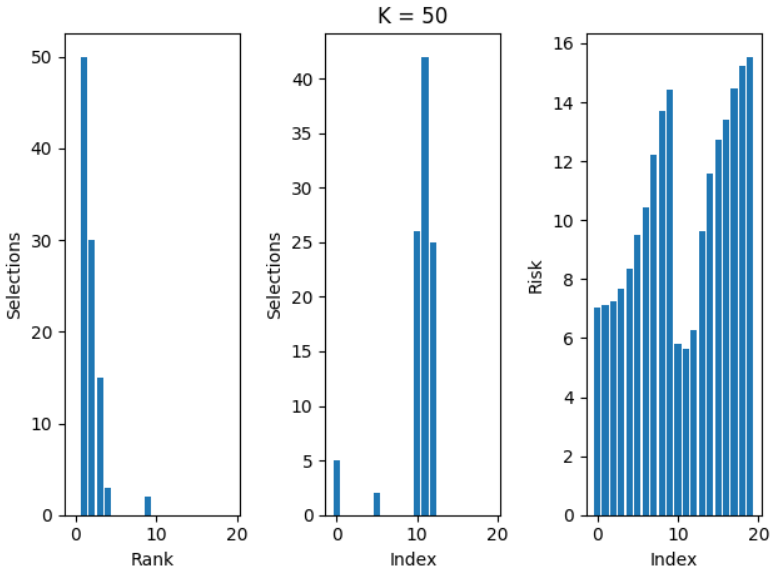

FIGURE 2 – $K = 50$.

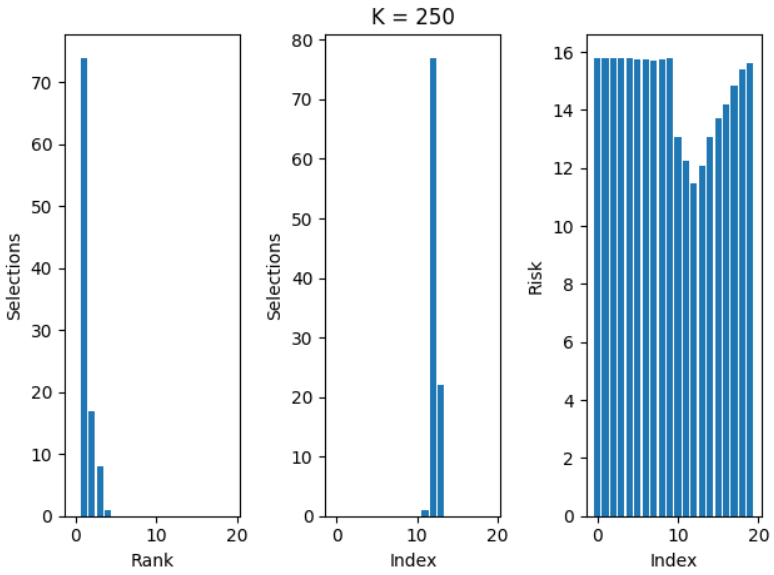

FIGURE 3 – $K = 250$.

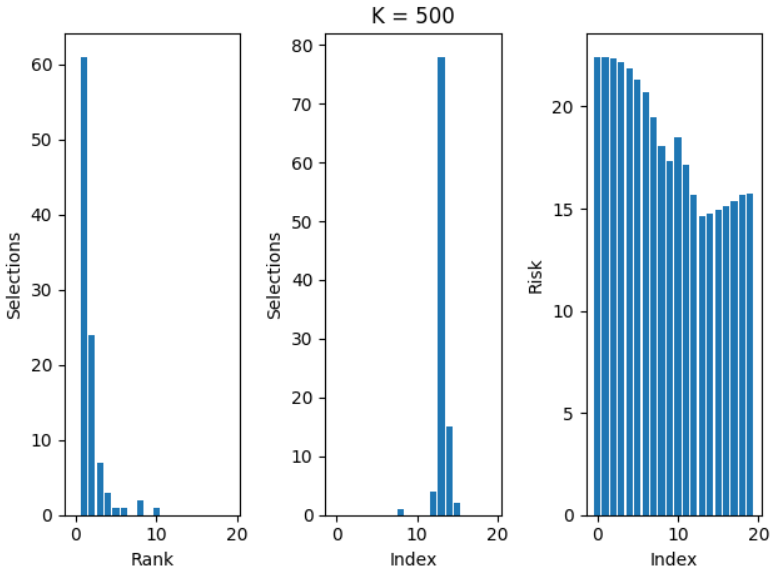

FIGURE 4 – $K = 500$.

On observe que l'estimateur choisi est souvent parmi les meilleurs. La procédure est relativement efficace dans tout les cas.

4 Conclusion

Dans un premier temps, nous avons introduit un estimateur naturel et universel. Nous avons établi une borne "intéressante" sur le risque de notre procédure de sélection, que nous pouvons spécifier dans divers cas particuliers. Nous avons ensuite vu que cette borne est en un sens "optimale" en établissant une borne minimax comme conséquence du lemme de Birgé. Nous avons enfin réalisé une application numérique dans le cas vectoriel, et mis en évidence la relative performance de notre critère.

Références

[1] Alexander Goldenshluger. A universal procedure for aggregating estimators. *The Annals of Statistics*, pages 542–568, 2009.