Introduction à l'estimation non paramétrique

October 11, 2020

1 Quelques notions générales de statistiques

1.1 Formalisme, notation et exemple

Pour formaliser une expérience statistique, il faut tout d'abord que l'observation soit issue d'un espace probabilisable $(\Omega, \mathcal{F})$. Sur cet espace, nous considérons un ensemble de lois de probabilités $\mathcal{P}$, qui contient l'ensemble de lois de probabilités que nous pensons susceptibles de régir l'expérience considérée. S'il est à priori possible de choisir pour $\mathcal{P}$ la totalité des lois de probabilité sur $(\Omega, \mathcal{F})$, il ne s'agit pas toutefois d'une obligation ; l'ensemble $\mathcal{P}$ permet de prendre en compte les informations connues à priori ; et ce choix est ainsi primordial.

L'espace $\mathcal{P}$ est généralement paramétré par un autre espace Θ, sous la forme $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$. Cette paramétrisation n'est pas toujours identifiable : il est possible que pour $\theta_1 \neq \theta_2$, $P_{\theta_1} = P_{\theta_2}$

Par souci de simplicité, nous considérons ici que nous observons un évènement $\omega \in \Omega$, tiré d'après une loi P. A priori, cette loi P peut ou peut ne pas appartenir à $\mathcal{P}$. L'objectif de la statistique est d'inférer de l'information sur la loi P à partir de l'observation ω. Cette information peut être partielle ou totale. Souvent, un statisticien cherche à construire un $\hat{\theta}$ tel que $P_{\hat{\theta}} \simeq P$. Encore faut il définir le sens de proximité sur un tel espace ! Mais il est possible qu'un statisticien cherche à déterminer une propriété plus restreinte de P (par exemple sa moyenne, sa variance), donc une fonction partant de $\mathcal{P}$ dans un autre espace. Nous nous concentrerons ici sur le cas de l'estimation de θ.

Usuellement, une expérience statistique est indexé par un nombre d'observations, noté n. Il est courant alors de supposer que la paramétrisation par Θ est indépendante de n, tandis que $\Omega, \mathcal{F}, \mathcal{P}$ le sont. Il s'agit alors de savoir, si, dans le cas où l'observation ω_n est tiré selon $P_{n,\theta}$ et quand le modèle est identifi-

able, nous avons $\lim_{n\to\infty} \hat{\theta}_n = \theta$, pour une convergence soit en probabilité ou presque sûre. Il faut donc que Θ soit muni au minimum d'une topologie.

L'objet $\hat{\theta}$ recherché est un estimateur. Un estimateur est une fonction mesurable de Ω dans Θ ; afin de le construire, aucune fonction de θ ne peut donc être employée. Un des objets des statisticiens est de construire un estimateur $\hat{\theta}$ ayant les propriétés de convergence présenté ci dessus ; dans le cas courant où de plus l'espace Θ est métrique, il s'agit de s'intéresser aux vitesses optimales d'estimation.

Donnons un exemple aisé pour rendre plus compréhensible nos notations : n lancers indépendant d'une pièce biaisée. Le paramètre déterminant l'expérience est la probabilité d'obtenir pile (noté 1), indexé par θ. L'espace $\Omega_n = \{0,1\}^n$, munie de la topologie discrète. Chacune des lois de probabilité $P_{n,\theta}$ correspond à n-lancers indépendants d'une pièce telle que la probabilité d'obtenir pile est θ. Notons donc que l'hypothèse d'indépendance des lancers est incorporée dans le choix de $\mathcal{P}$. Ici, $\Theta = [0,1]$ est métrique, et est doté de la topologie découlant de cette métrique. Un estimateur assez évident de θ est la proportion de pile observé. Cet estimateur converge presque surement vers θ quand n tend vers l'infini.

Le formalisme développé ci-dessus permet de traiter des cas bien plus complexes, où l'espace θ n'est pas une partie d'un espace de dimension fini. Il peut s'agir par exemple d'un espace fonctionnel ; par exemple des paramètres d'un processus de diffusion de type $dX_t = \sigma(X_t)dB_t$ où B_t est un mouvement brownien, pour l'observation de $(X_1, \ldots, X_n)$. L'estimation de σ est un exemple d'estimation non paramétrique.

1.2 Vitesse d'estimation

Si l'estimation n'est pas le seul objet de la statistique, cette présentation se restreindra à celui-ci. Plus précisément, nous nous intéressons au cas où l'objet à estimer appartient à un espace métrique. Grossièrement, la vitesse d'estimation d'un estimateur représente la vitesse à laquelle la suite $(\hat{\theta}_n - \theta)_{n\in\mathbb{N}}$ tend vers 0. La vitesse d'estimation n'est pas une valeur ponctuelle, mais globale sur Θ. Pour un problème statistique donné, il s'agit par exemple de montrer que

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \; d(\hat{\theta}_n, \theta) \leq r_n$$

Pour une suite r_n tendant vers 0, qui donne une majoration de la vitesse d'estimation. D'autres critères peuvent toutefois être avancés, par exemple : pour tout θ, avec probabilité au moins $\alpha_n \to_{n\to\infty} 1$,

$$d(\hat{\theta}_n, \theta) \leq r_n$$

Ou même, pour $\Theta \subset \mathbb{R}^n$, $\forall \theta$,

$$r_n^{-1}(\theta - \hat{\theta}_n) \to_{\mathcal{D}} \mathcal{L}$$

pour une certaine loi $\mathcal{L}$.

Ce type de résultat permet d'établir une borne supérieure sur la vitesse d'estimation. Pour que l'estimateur soit jugé bon, il faut s'assurer que sa vitesse n'est pas trop éloignée de la vitesse optimale qu'un estimateur peut atteindre. Dans le paradigme minimax, le risque minimax est défini par

$$R_n = \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta d(\hat{\theta}_n, \theta)$$

où la minimisation est faite sur tous les estimateurs possibles. Une famille d'estimateurs $(\hat{\theta}_n)_{n\in\mathbb{N}}$ réalisant ce risque à une constante près est dit minimax ; si de plus, $R_n^{-1} \sup_{\theta\in\Theta} \mathbb{E}_\theta \; d(\hat{\theta}_n, \theta) \to 1$, l'estimateur est sharp.

Il est clair que la vitesse d'estimation dépend de la distance d choisie. Dans un cadre plus général, il n'est pas nécessaire, ni toujours judicieux, de demander que d soit une distance ; on parle alors de fonction de perte, fonction à valeur dans $\mathbb{R}^+$ tel que $\forall \theta$, $L(\theta, \theta) = 0$. La même approche minimax reste applicable.

1.3 Estimation paramétrique

Dans le cadre d'estimation paramétrique, l'objet de l'estimation vis dans une sous partie de $\mathbb{R}^d$. Un des résultats les mieux connus d'estimation paramétriques s'applique dans le cas d'expériences produits, c'est à dire où une même expérience est répétée de manière indépendante n fois, avec quelques hypothèses supplémentaires. La vitesse d'estimation "attendue" dans le cadre de l'estimation paramétrique est de l'ordre de $n^{-\frac{1}{2}}$, à constante près. Si cette vitesse d'estimation n'est pas automatique, son apparition est établie dans un cadre plus restreint par deux résultats :

Theorem 1.1. *Dans un modèle exponentiel canonique minimal, l'estimateur $\hat{\theta}_n$ dit du maximum de vraisemblance est asymptotiquement normal :*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to_{\mathcal{D}} \mathcal{N}\left(0, I(\theta)^{-1}\right)$$

où $I(\theta)$ est la matrice d'information de Fischer (échantillonné une fois).

Theorem 1.2. *Théorème de Cramer–Rao*

Dans le cas de n observations indépendantes, pour $\hat{\theta}$ un estimateur sans biais de θ,

$$\mathbb{E}_\theta \left\|\hat{\theta} - \theta\right\|^2 \geq \frac{1}{n} \left\|I(\theta)^{-1}\right\|_2^2$$

où $I(\theta$ est la matrice d'information de Fischer (échantillonné une fois).

Rappelons toutefois que cette vitesse paramétrique n'est pas systématique. Il est par exemple assez aisé d'en donner un contre exemple : l'estimation du maximum m atteignable par une variable aléatoire uniforme sur $[0, m]$ est plus rapide, de vitesse n^{-1}. De plus, la borne de Cramer–Rao ne porte que sur des estimateurs sans biais ; il est à priori possible de trouver un estimateur biaisé meilleur.

Un grand nombre de modèles rentrant dans le cadre des modèles exponentiels : gaussiennes multivariées, lois de poissons, lois gammas... expliquant l'intuition assez générale qu'un problème statistique paramétrique standard doit avoir une vitesse d'estimation de l'ordre de $\sqrt{\frac{d}{n}}$, où d est la dimension. Le cas où la dimension croit avec n (par exemple $d = n$) relève de la statistique en grande dimension ; il est alors nécessaire, afin de pouvoir construire un estimateur consistent, d'ajouter d'autre contraintes : un exemple connu est celui de l'estimation de matrices à partir d'observations tronquées (cas où par exemple $n \ll d$) en supposant une structure de faible rang (voir [17], [20])

De même, l'estimation non paramétrique nécessite des hypothèses supplémentaires pour fonctionner. Ces contraintes peuvent être sous la forme d'hypothèses de régularité, de contraintes de formes ou autres. Il est en tout cas attendu en estimation non paramétrique d'obtenir des vitesses plus lentes que les vitesses paramétriques.

2 Estimation non–paramétrique

L'estimation non–paramétrique est un domaine vaste, qui apparaît naturellement dès qu'il s'agit d'estimer un objet sans hypothèse forte sur celui-ci, et permet, au prix de méthodes plus complexes, d'inférer des résultats bien plus proches de la réalité. Il est en effet illusoire de s'imaginer que tous les phénomènes présents dans la nature s'approximent par des phénomènes gaussiens. Nous donnons quelques exemples d'estimation non–paramétrique, qui permettent d'obtenir une bonne intuition sur le rôle que jouent la dimension ambiante ainsi que la régularité.

2.1 Estimation de densité

L'estimation de densité, sans à priori sur l'appartenance de celle ci à une grande famille de lois (gaussiennes multivariées,), est un champ naturel de l'estimation non–paramétrique, et permet de comprendre quelques grands principes. Nous supposons observer des tirages indépendants d'une même loi P sur $E \subset \mathbb{R}^d$. L'objet de l'estimation non paramétrique est d'inférer la

loi de P, qui est supposée avoir une densité p par rapport à Lebesgue. La distance considérée sera ici une norme $L^q(E)$.

Donnons maintenant deux résultats qui peuvent de prime abord sembler contradictoire, montrant la nécessité d'ajouter d'autres hypothèses avant de pouvoir espérer trouver des vitesses de convergence :

Theorem 2.1. *Il est possible de construire un estimateur à noyau $\hat{p}_n$ consistent pour le risque L^1, c'est à dire tel que $\forall p$,*

$$\lim_{n\to\infty} \mathbb{E} \int_{x\in E} |\hat{p}_n(x) - p(x)| = 0$$

Theorem 2.2. *Il n'existe aucune vitesse de convergence pour l'estimation d'une densité sur E (sous réserve que E contienne un ouvert de $\mathbb{R}^d$). Formellement,*

$$\sup_p \inf_{\hat{p}_n} \mathbb{E} \int_{x\in E} |p(x) - \hat{p}_n(x)| = 2$$

Ce qui signifie que, bien que nous ayons établi l'existence d'un estimateur consistent, la vitesse d'estimation peut être arbitrairement lente, mettant à mal le paradigme minimax. Pour contrer cette difficulté, il faut ajouter une hypothèse de régularité sur les densités p à estimer. Cette régularité est généralement quantifié par un paramètre s. Il peut s'agir d'espaces de Sobolev ou de Besov, ou de fonctions Höldériennes. s correspond alors au niveau de dérivabilité de la densité. Bien que cette propriété ne puisse être établie sans d'autres précisions, il est possible de montrer dans certains cas que la vitesse d'estimation est alors de $n^{-\frac{s}{2s+d}}$. La construction d'un estimateur atteignant cette vitesse peut se faire par une méthode de noyau suffisamment régulier, décrite ci-dessous.

Soit K un noyau, c'est à dire une fonction intégrable de $\mathbb{R}^d$ d'intégrale 1 ; on appelle K_h la fonction définie par $K_h(x) = h^{-d}K(\frac{x}{h})$. Nous définissons l'estimateur $\hat{f}_h$ par $\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) * K_h$, où $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ est le processus empirique. Il est clair que $\mathbb{E}(\hat{f}_h) = f * K_h$, et que nous avons donc affaire à un estimateur biaisé. Nous procédons à une décomposition type biais–fluctuation :

$$\mathbb{E}_f \left\| f - \hat{f} \right\|_{L^p} \leq \| f - f * K_h \|_{L^p} + \mathbb{E}_f \left\| f * K_h - \hat{f} \right\|_{L^p}$$

Il s'agit alors de choisir h de manière à ce que les deux termes soient de même ordre, et de préciser des conditions sur K. Selon la régularité s de f, l'action de la convolution par K sera plus ou moins importante. Intuitivement, la convolution a un caractère lissant, et si la fonction f oscille à petite échelle, ces oscillations seront gommés, résultant dans un terme de biais plus

important. Au contraire, plus f est régulier, moins la densité sera perturbée par la convolution avec K_h. En choisissant K de manière appropriée (ce qui dépend de s), et en supposant qu'on ait un contrôle uniforme sur une norme dépendant de la régularité s des f possibles (par exemple une condition de type $\|f\|_{\mathcal{H}^s} \leq C$, on peut obtenir une borne $\|f - f * K_h\|_{L^p} \leq Ch^s$. Le terme de fluctuation peut être borné dans un certain nombre de cas par $C\sqrt{\frac{1}{h^d n}}$. Afin de minimiser notre borne, il suffit de prendre $h \simeq n^{-\frac{1}{2s+d}}$: l'estimateur ainsi construit a une vitesse majorée par $n^{-\frac{s}{2s+d}}$.

Le procédé présenté ici passe sous silence différentes conditions nécessaires afin de mener à bien l'analyse. Elles peuvent s'exprimer sous la forme de condition de support compact au diamètre majoré par une constante connue, d'une décroissance à vitesse fixée à l'infini... Travailler sur un tore permet souvent de simplifier l'analyse, enlevant tout problème de bord et permettant tout de même de convoluer facilement. Il n'existe pas, à la connaissance de l'auteur, de résultat général sur les méthodes d'estimation à noyau : [10] en établit un pour des densités appartenant à des espaces de Besov $B^s_{p\infty,}$, et [34] donne un résultat pour $d = 1$, des densités dans des classes de Holder ou de Sobolev et des pertes quadratiques, locale pour Holder et intégrée pour Sobolev (voir aussi [8] et [9] pour une vue des résultats quand $p = 1$). Si d'autres méthodes existent (par projection notamment), les estimateurs à noyaux ont l'avantage de la simplicité, une fois les noyaux réguliers établis (voir [34] pour la construction).

2.2 Équations de diffusions

La recherche des dynamiques bio-moléculaires, l'étude de phénomènes économiques créent un besoin d'estimer des processus de manière fine, et en particulier sans hypothèse paramétrique. Une sous classe de processus particulièrement étudiée est les processus de diffusion (voir [1], [4], [7], [11], [16], [28], [13], [35], [26]) Les diffusions considérées sont des diffusions d'Ito sur $\mathbb{R}^d$, solutions de l'équation différentielle stochastique

$$dY_t = b(Y_t)dt + \sigma(Y_t)dW_t. \tag{1}$$

L'objectif est d'estimer le processus à partir d'observations discrètes espacées d'un pas de temps Δ : c'est à dire qu'on observe $(Y_0, Y_\Delta, \ldots, Y_{n\Delta})$. Plusieurs cas de figure sont envisageables ; par exemple $\Delta \to_{n\to\infty} 0$ (régime de haute fréquence), soit le pas de temps est fixe (régime de basse fréquence). Le pas de temps peut être aléatoire, observé ou non. Par la suite, nous considérons que le pas de temps est fixé, et vaut 1. Le résultat qui suit permet

d'introduire les méthodes par projection, autre instrument utile en estimation non paramétrique.

Cette diffusion est aussi décrite par les probabilités de transition $p_x(\cdot) = p(x,\cdot)$. L'estimation de la probabilité de transition consiste à l'estimation de la densité du couple (Y_0, Y_Δ), ce qui revient peu ou prou à ce qui est décrit précédemment dans le cadre d'estimation de densité ; la vitesse d'estimation est donc, en supposant que p appartient à l'espace de Sobolev H^s, $n^{-\frac{s}{2s+2d}}$ (l'espace ambiant étant de dimension $2d$, nous trouvons la vitesse d'estimation attendue ; et remarquons au passage que l'hypothèse d'indépendance des observations peut être affaiblie). Pour une preuve de ce résultat, voir [5], [22], [23].

Nous allons par la suite montrer les grandes étapes d'une amélioration de ce résultat, en utilisant la structure particulière qu'offre la diffusion. Les paramètres b et σ de la diffusion définissent de manière injective les probabilités de transition. Celles ci sont de même entièrement déterminées par l'opérateur de transition, noté P, et nous admettrons que la norme L^2 sur p équivaut à la norme de Hilbert–Schmidt sur P. En supposant la diffusion réversible, cet opérateur est elliptique, auto adjoint pour la mesure invariante μ de la diffusion, et a une décomposition spectrale dont la décroissance des valeurs propres est exponentielle, d'après la loi de Weyl.

Afin d'éviter les problèmes de bord, nous considérons une diffusion sur le tore $\mathbb{T}^d$, décrite par l'équation (1). Nous demandons, pour simplifier les preuves, que la mesure invariante de cette diffusion soit la mesure de Lebesgue.

Pour $(\Psi_i)_{i\in\mathbb{N}}$ une base orthonormée de $L^2(\mathbb{T}^d)$, l'opérateur de transition P est déterminé par l'ensemble des coefficients $(\langle \Psi_i, P\Psi_{j'}\rangle)_{(i,j)\in\mathbb{N}^2}$. Afin d'obtenir un estimateur de P, nous allons estimer une partie de ces coefficients par $\hat{P}_{i,j}$ pour $i,j \le N$ et considérer l'opérateur défini par $\hat{P}_{i,j} = \mathbb{1}(i,j \le N)\hat{P}_{i,j}$. L'erreur d'estimation se scinde ensuite en deux parties : un terme de biais $\sum_{\max i,j>N} P_{i,j}^2$ et un terme de fluctuation $\sum_{i,j\le N}(P_{i,j} - \hat{P}_{i,j})^2$. Un choix approprié de N permettra finalement de balancer fluctuation et biais. On note P_N la réduction de P à $V_N = \mathrm{Span}(\Psi_i \mid i \le N)$, c'est à dire, en notant π_N la projection orthogonale sur V_N, $P_N = \pi_N P \pi_N$. La séparation biais fluctuation découle de la séparation $P - \hat{P} = (P - P_N) + (P_N - \hat{P})$. Cette approche de construction d'estimateur s'appelle l'estimation par projection.

L'estimation de $(P_{i,j})_{i,j\le N}$ se fait en deux étapes : tout d'abord en construisant un premier estimateur $\tilde{P}$ qui estime individuellement chaque composante, puis en utilisant la propriété de décroissance des valeurs propres de

P pour réduire le terme de fluctuation. On définit

$$\tilde{P}_{i,j} = \frac{1}{2n} \sum_{0 \leq m \leq n-1} (\psi_i(Y_m)\psi_j(Y_{m+1}) + \psi_j(Y_m)\psi_i(Y_{m+1}))$$

Cet estimateur a de bonnes propriétés d'approximation de P_N en norme d'opérateur ; avec grande probabilité, $\left\|\tilde{P} - P_N\right\|_\infty \leq \sqrt{\frac{N}{n}}$. Pour transformer cette borne en norme d'opérateur en borne sur la norme d'Hilbert-Schmidt, nous utilisons le fait que P, et donc P_N aussi, s'approxime bien par un opérateur de rang faible, ce qui découle de la décroissance exponentielle de ses valeurs propres. Rappelons que $\|A\|_2 \leq \mathrm{Rang}(A)\,\|A\|_\infty$; avoir une structure de faible rang permet donc d'obtenir une bonne borne sur la norme d'Hilbert–Schmidt.

Une technique fréquemment utilisée pour imposer à un estimateur une structure de faible rang (ou de parcimonie) est le seuillage. Ici, la matrice $\tilde{P}$ est symétrique, et donc diagonalisable sous forme $\hat{Q}\hat{D}\hat{Q}^t$. Le seuillage de $\tilde{P}$ se fait sur ses valeurs propres, en replaçant par 0 toutes celles dont la valeur absolue est en dessous d'un seuil α. Il peut aussi se définir de manière équivalente par

$$\hat{P} = \arg\min_S (\left\|\tilde{P} - S\right\|^2 + \alpha^2 \mathrm{Rang}(S)) \tag{2}$$

Il est facile de montrer que pour $\alpha \geq \left\|\tilde{P} - P_N\right\|_\infty$ cet estimateur satisfait la propriété suivante :

$$\left\|\hat{P} - P_N\right\|_2^2 \leq 4 \inf_S (\|S - P_N\|_2^2 + \alpha^2 \mathrm{Rang}(S)) \tag{3}$$

En choisissant pour S une bonne approximation de rang fini de P_N, on obtient avec grande probabilité la borne $\left\|\hat{P} - P_N\right\|_2^2 \leq \frac{N}{n} \log(n)^{\frac{d}{2}}$.

Afin de borner le terme de biais, nous supposons, ce qui est justifié dans le cadre de diffusion qui nous intéresse, que l'opérateur P se décompose dans une base orthogonale $(u_k)_{k \in \mathbb{N}}$ qui appartiennent à l'espace de Sobolev H_s. On demande de plus à ce que $\sum_k \lambda_k^2 \left\|u_k\right\|_{H_s}^2 < C$. Le terme de biais se décompose en

$$\|\pi P \pi - P\|_2 \leq \|\pi P \pi - \pi P\|_2 + \|\pi P - P\|_2 \leq \|P\pi - P\|_2 + \|\pi P - P\|_2 \leq 2\,\|(Id - \pi)P\|_2$$

Utilisons maintenant l'identité

$$\begin{aligned} \|P(Id - \pi)\|_2^2 &= \sum_k \|(Id - \pi)Pu_k\|_2^2 \\ &= \sum_k \lambda_k^2 \|(Id - \pi)u_k\|_2^2 \end{aligned}$$

Jusqu'ici, nous n'avons pas eu besoin d'hypothèse sur la régularité de la base Ψ choisie. L'identité précédente montre qu'une base permettant de bien estimer une fonction suffisamment lisse avec un nombre restreint d'éléments permettrait d'obtenir une bonne borne. Nous choisissons donc d'employer une base d'ondelettes Ψ_λ de régularité $S > s$ et en considérant pour V_N l'ensemble des ondelettes de résolution au maximum de J. Nous avons donc $N = \dim(V_N) \propto 2^{Jd}$. Les résultats connus sur les ondelettes entraînent que

$$\|(Id - \pi)u_k\|_2 \leq 2^{-Js} \|u_k\|_{H_s}$$

D'où il découle que le biais est borné par 2^{-Js}. En réunissant les 2 bornes, nous obtenons une majoration de l'erreur d'estimation avec grande probabilité de $2^{-Js} + 2^{\frac{Jd}{2}} n^{-\frac{1}{2}} \log(n)^{\frac{d}{4}}$. En optimisant sur J, la vitesse d'estimation est majorée par $Cn^{-\frac{s}{2s+d}} \log(n)^{-\frac{d}{2}\frac{s}{2s+d}}$, ce qui, au facteur logarithmique près, est la vitesse d'estimation non paramétrique avec cette fois ci une dimension de d plutôt que $2d$. On peut montrer que le facteur logarithmique ne peut pas être enlevé, et que cet estimateur est minimax optimal. Pour une preuve détaillée de ce résultat dans un cadre plus général, voir [25].

3 Bornes inférieures

La recherche de la vitesse minimax d'estimation se fait en deux étapes : la construction d'un estimateur atteignant cette vitesse, et la preuve qu'aucun autre estimateur ne peut être plus rapide. Cette deuxième partie s'intéresse à l'information contenue dans l'observation, c'est à dire à la géométrie de l'espace des lois de probabilités considérées. Intuitivement, si les lois sont paramétrées par θ, nous cherchons à savoir comment $P_{\theta'}$ s'approche de P_θ pour $\theta' \to \theta$; plus les lois s'approchent lentement, plus l'estimation autour de θ est facile, et donc plus la borne inférieure est faible. A l'inverse, si on peut trouver deux lois similaires P_θ et $P_{\theta'}$ tels que $d(\theta, \theta')$ est grand[1], on peut obtenir une borne inférieure assez grande. L'obtention de bornes inférieures se rapproche donc de questions de tests statistiques.

Un test T est une statistique à valeur dans $\{0, 1\}$ cherchant à établir ou rejeter une hypothèse contre une autre. Dans le cas d'hypothèses simples, il s'agit de savoir si une loi est tirée selon une loi P_0 ou une loi P_1. L'erreur de première espèce consiste à rejeter $X \simeq P_0$ sous P_0, l'erreur de seconde espèce à accepter $X \simeq P_0$ sous P_1. Le niveau d'un test est dans ce cadre $\mathbb{E}_{P_0}(T)$,

[1]Notons que la notion de distance entre lois de probabilité est ambigüe, plusieurs distances naturelles existant ; les plus utiles sont la distance d'Hellinger, en variation totale, et la divergence de Kullback–Leibler, qui n'est toutefois pas une distance

et sa puissance $\mathbb{E}_{P_1}(T)$. Le lemme fondamental de Neyman–Pearson indique qu'un test au maximum de vraisemblance ont une puissance maximale parmi les tests de niveau inférieur ou égal.

Une méthode habituelle pour établir des bornes inférieures repose sur les remarques suivantes :

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left(d(\theta,\hat{\theta})\right) &\geq s \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta\left(d(\theta,\hat{\theta}) \geq s\right) \\ &\geq s \inf_{\hat{\theta}} \sup_{\theta\in\{\theta_0,\ldots,\theta_M\}} \mathbb{P}_\theta\left(d(\theta,\hat{\theta}) \geq s\right)\end{aligned}$$

Si s est de l'ordre de la vitesse r_n que l'on cherche à établir comme borne inférieure de la vitesse d'estimation, il suffit donc de borner $\inf_{\hat{\theta}} \sup_{\theta\in\{\theta_0,\ldots,\theta_M\}} \mathbb{P}_\theta\left(d(\theta,\hat{\theta}) \geq s\right)$ par une constante. En supposant maintenant que l'ensemble $\{\theta_0,\ldots,\theta_M\}$ est constitué d'éléments suffisamment éloignés les uns des autres, c'est à dire tel que $\forall i \neq j$, $d(\theta_i,\theta_j) \geq 2s$ on peut restreindre l'infimum pris sur $\hat{\theta}$ à un infimum sur les estimateurs à valeur dans $\{\theta_0,\ldots,\theta_M\}$, ce qui revient à écrire chercher à borner, sur les estimateurs Ψ à valeur dans $[0,M]$, $\inf_\Psi \sup_j \mathbb{P}_{\theta_j}(\Psi \neq j)$. Nous terminons cette brève présentation des bornes inférieures avec un résultat central :

Theorem 3.1. *S'il existe des $\theta_0,\ldots,\theta_M$, tels que les mesures P_i soient toutes absolument continues les unes entre elles*[2]*, tels que $\forall i \neq j$ $d(\theta_i,\theta_j) \geq 2s$, et s'il existe $\tau > 0$ et $0 < \alpha < 1$ tels que*

$$\frac{1}{M}\sum_{i=1}^{M} P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) \geq 1-\alpha \tag{4}$$

alors

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} P_\theta(d(\theta,\hat{\theta} \geq s) \geq \frac{\tau M}{1+\tau M}(1-\alpha) \tag{5}$$

La condition (4) se décline de plusieurs façons pour donner des conditions sur les distances des différentes lois entre elles, pour plusieurs distances possibles.

4 Conclusion

L'estimation non paramétrique forme un domaine de recherche actif en statistiques, et relativement neuf ; si longtemps des statistiques paramétriques

[2]ceci peut être relâché, il suffit de considérer la partie absolument continue de P_0 par rapport à P_j dans la formule qui suit

ont été préféré pour leur implémentation plus aisé, l'augmentation des puissances de calcul ainsi que la prise de conscience des limites des hypothèses paramétriques a poussé à un gain d'intérêt pour ce domaine. Plusieurs questions demeurent encore à résoudre ; tout d'abord la mise en place d'un modèle général permettant de rassembler la plupart des intuitions des statisticiens sur les modèles non paramétriques, l'obtention d'estimateurs ayant des garantie de robustesse (que se passe-t-il quand la loi P selon laquelle les données sont tirées n'appartient pas au modèle?) ou d'adaptation (peut on construire un adaptateur qui ne nécessite pas la connaissance préalable de s, tout en atteignant la vitesse minimax). Une dernière question soulevée depuis quelques années concerne la prise en compte des contraintes algorithmiques : peut on établir des résultats sur des vitesses minimax en considérant des estimateurs dont la complexité algorithmique est limitée (voir [2])?

References

[1] K. Abraham. Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data. *Bernoulli, to appear*, 2018.

[2] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[3] L. Birgé. *Robust tests for model selection. From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, pages 47–64. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2013.

[4] Jakub Chorowski and Mathias Trabs. Spectral estimation for diffusions with random sampling times. *Stochastic processes and their applications*, 126(10):2976–3008, 2016.

[5] S. Clemençon. *Methodes d'ondelettes pour la statistique non parametrique des chaines de Markov*. PhD thesis, Universite Paris 7, 2000.

[6] R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

[7] A. Dalalyan. Sharp adaptative estimation of the drift function for ergodic diffusions. *Ann. Statist.*, 33(6):2507–2528, 2005.

[8] L Devroye and L Gyrfi. Nonparametric density estimation: the l1 view. john wiely & sons, 1985.

[9] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[10] E. Gine and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[11] Emmanuel Gobet, Marc Hoffmann, Markus Reiß, et al. Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, 32(5):2223–2253, 2004.

[12] Lars Peter Hansen, José Alexandre Scheinkman, and Nizar Touzi. Spectral methods for identifying scalar diffusions. *Journal of Econometrics*, 86(1):1–32, 1998.

[13] M. Hoffmann. Adaptive estimation in diffusion processes. *Stochastic Process. Appl.*, 79(1):135–163, 1999.

[14] L. Hörmander. The Weyl Calculus of Pseudo-Differential Operators. *Comm. Pure Appl. Math.*, 32:359–443, 1979.

[15] V. Ivrii. Sharp spectral asymptotics for operators with irregular coefficients. *Int. Math. Res. Notices*, 2000(22):1155–1166, 2000.

[16] Mathieu Kessler, Michael Sørensen, et al. Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(2):299–314, 1999.

[17] O. Klopp. Noisy low-rank completion with general sampling distribution. *Bernoulli*, 2014.

[18] V. Koltchinskii and K. Lounici. Asymptotics and Concentration Bounds for Bilinear Forms of Spectral Projectors of Sample Covariance. *Ann. Henri PoincarÃ©*, 52(4):1976–2013, 2016.

[19] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy Low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[20] Vladimir Koltchinskii and Dong Xia. Optimal Estimation of Low Rank Density Matrices. *J Mach Learn Res.*, 16:1757–1792, 2015.

[21] Y. A. Kutoyants. On nonparametric estimation of trend coefficient in a diffusion process. *Statistics and Control of Stochastic Processes*, pages 230–250, 1984.

[22] C. Lacour. Adaptative estimation of the transition density of a Markov Chain. *Ann. Henri PoincarÃ©*, 2007.

[23] C. Lacour. Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Process. Appl.*, 118(2):232–260, 2008.

[24] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[25] Matthias Löffler and Antoine Picard. Spectral thresholding for the estimation of markov chain transition operators, 2020.

[26] Grigori N. Milstein, John G. M. Schoenmakers, and Vladimir Spokoiny. Transition density estimation for stochastic differential equations via foraward-reverse representations. *Bernoulli*, 10, 2004.

[27] R. Nickl and K. Ray. Nonparametric statistical inference for diffusion processes with high-dimensional Gaussian priors. *unpublished manuscript*, 2018.

[28] R. Nickl and J. Söhl. Nonparametric bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.*, 45(4):1664–1693, 2017.

[29] D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20(79):32, 2015.

[30] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134:124116, 2011.

[31] C. Strauch. Sharp adaptive drift estimation for ergodic diffusions: the multivariate case. *Stochastic Process. Appl.*, 125(7):2562–2602, 2015.

[32] C. Strauch. Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions. *Probab. Theory Related Fields*, 164(1):361–400, 2016.

[33] C. Strauch. Adaptive invariant density estimation for ergodic diffusions over anisotropic classes. *Ann. Statist., to appear*, 2018.

[34] A.B. Tsybakov. *Introduction to nonparametric estimation.* Springer, 2008.

[35] J. van Waaij and H. van Zanten. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electron. J. Statist.*, 2016.

[36] H. Weyl. Über die Asymptotische Verteilung der Eigenwerte. *Nachr. Königl. Ges. Wiss. Göttingen*, pages 110–117, 1911.

[37] A. Zhang and M. Wang. State Compression of Markov Processes via Empirical Low-Rank Estimation. *ArXiv preprint*, 2018.