Modèles de mélange semi-paramétriques et applications aux tests multiples. Document de synthèse.

Van Hanh Nguyen

Dans cette synthèse, nous allons tout d'abord présenter le contexte de ce travail de thèse, à savoir la problématique des tests multiples, les différentes mesures de l'erreur des procédures de test multiple ainsi que le point de vue de la statistique semi-paramétrique sur ces problèmes. Cette introduction emprunte des éléments et points de vue à [12, 14, 16, 19]. Nous détaillerons ensuite les deux contributions majeures de cette thèse, développées dans les chapitres 2 et 3 du manuscrit.

1 Contexte statistique de la thèse

1.1 Tests multiples

La problématique des tests multiples est relativement ancienne en statistique, mais elle a été remise au goût du jour au début des années 2000 avec l'arrivée de nouvelles données dites « de grande dimension ». De nouvelles applications sont apparues comme par exemple l'analyse de « puces » en génomique [3], l'astrophysique [8] ou la neuro-imagerie [18].

Rappelons tout d'abord le contexte des tests statistiques (simples). On souhaite tester une hypothèse H_0 (que l'on pourra noter de façon équivalente $H = 0$) contre une alternative H_1 (ou encore $H = 1$) au vu de l'observation (ou statistique) X. Pour une région de rejet donnée Γ, l'hypothèse H_0 est rejetée lorsque $X \in \Gamma$ et acceptée sinon. Une erreur de type I a lieu lorsque l'hypothèse nulle (H_0) est à la fois vraie et rejetée tandis qu'une erreur de type II apparaît lorsque l'hypothèse nulle est fausse et acceptée. Afin de choisir cette région de rejet Γ, un taux d'erreur de type I maximum $\alpha > 0$ est fixé et on considère toutes les régions de rejet dont l'erreur de type I est inférieure à ce niveau α. Parmi toutes ces régions, celle(s) qui possède(nt) la plus petite erreur de type II est alors sélectionnée. Ainsi, la région de rejet est construite afin de contrôler une erreur de type I. Plus précisément, on cherche une région de puissance maximale (puissance = 1- erreur de type II) en contrôlant au niveau α l'erreur de type I.

Dans le cas des tests multiples, la situation est plus complexe. Considérons par exemple le cas où on souhaite tester simultanément $n = 10000$ hypothèses, parmi lesquelles $n_0 = 8000$ vérifient que H_0 est vraie, au niveau $\alpha = 5\%$ pour chaque test. Alors la procédure utilisée fait en moyenne $n_0\alpha = 400$ faux positifs (ou erreurs de type I), ce qui semble inadapté. Une procédure de test multiple vise à contrôler a priori le niveau de chaque test afin d'obtenir une certaine « quantité » de faux positifs sous un niveau fixé α. Cette quantité de faux positifs est mesurée à travers des taux globaux de faux positifs, comme par

exemple la probabilité de faire au moins une erreur de type I parmi toutes les hypothèses (FWER pour « family-wise error rate ») ou la proportion attendue de faux positifs parmi les hypothèses rejetées (FDR pour « false discovery rate »).

Avant de conclure cette introduction, on rappelle la définition de deux quantités fondamentales : la p-valeur et la z-valeur. Le degré de significativité ou p-valeur associée à un test est la probabilité sous l'hypothèse H_0, d'observer une valeur au moins aussi extrême que la statistique effectivement observée. En d'autres termes, la p-valeur est la plus petite probabilité sous H_0 que la statistique X appartienne à une région de rejet, parmi l'ensemble des régions de rejet emboîtées qui contiennent la statistique de test. Formellement, on peut écrire la p-valeur correspondant à l'observation $X = x$ sous la forme

$$p\text{-valeur}(x) = \inf_{\{\Gamma; x \in \Gamma\}} \mathbb{P}(X \in \Gamma | H_0),$$

où $\{\Gamma; x \in \Gamma\}$ est un ensemble de régions de rejet emboîtées qui contiennent la valeur observée x. On peut noter que toute p-valeur est, sous la loi H_0, stochastiquement majorée par une variable de loi uniforme, i.e.

$$\mathbb{P}(p(X) \leq t | H_0) \leq t, \text{ pour tout } t \in [0,1]. \tag{1}$$

Si la loi de la statistique X est absolument continue, alors l'égalité est vraie dans (1), i.e. la p-valeur suit, sous la loi H_0, une loi uniforme sur $[0,1]$.

Remarque 1. *Une région de rejet portant sur la p-valeur est toujours de la forme* $[0,\gamma]$ *pour un certain* $\gamma > 0$. *En effet, pour toutes p-valeurs* $p_1 \leq p_2$, *quelles que soient les valeurs des statistiques* x_1 *et* x_2 *associées, d'après la définition d'une p-valeur on a :* $x_2 \in \Gamma$ *implique* $x_1 \in \Gamma$. *En conséquence, si* p_2 *est rejetée alors* p_1 *l'est aussi.*

Enfin, la z-valeur est définie via la transformation probit

$$Z = \text{probit}(P) = \Phi^{-1}(P),$$

où P est la p-valeur et Φ la fonction de répartition de la loi normale centrée réduite.

1.2 Modèle de mélange et test multiple

On suppose à présent que l'on teste n hypothèses identiques $H_1, \ldots, H_n$ à partir des statistiques de test respectives $X_1, \ldots, X_n$. Par tests identiques on entend que le même type de région de rejet est utilisée pour chaque test. On pose $H_i = 0$ lorsque la $i$ème hypothèse nulle est vraie et $H_i = 1$ sinon. On note également $T_i = T(X_i)$ une transformation de la statistique de test X_i, par exemple la p-valeur P_i, la z-valeur Z_i, le taux local de faux positifs (défini ci-après) $\ell\text{FDR}(X_i)$, ou bien encore la statistique de test elle-même X_i. On suppose que les variables T_i sous l'hypothèse nulle $H_i = 0$ (resp. alternative $H_i = 1$) sont i.i.d. de fonction de répartition G_0 connue (resp. G_1 inconnue). Enfin, on suppose que les variables $H_i \in \{0,1\}$ sont i.i.d. de loi de Bernoulli de paramètre $\theta = \mathbb{P}(H_i = 0)$ inconnu. La loi marginale de T_i est alors donnée par un modèle de mélange à deux composantes

$$G(x) = \theta G_0(x) + (1-\theta) G_1(x),$$

et on note $g = \theta g_0 + (1-\theta) g_1$ les densités associées (lorsqu'elles existent). Lorsque les statistiques de test X_i sont des variables de loi continue sous l'hypothèse nulle $H_i = 0$, les

	H^i Acceptés	H^i Rejetés	Total
H^i vraie	TN	FP	n_0
H^i fausse	FN	TP	n_1
Total	W	R	n

TABLE 1 – Résultats possibles à partir d'une procédure de test de n hypothèses $H^1, \ldots, H^n$

p-valeurs associées P_i suivent la loi uniforme $\mathcal{U}[0,1]$ sous l'hypothèse nulle $H_i = 0$. De plus, la loi marginale des p-valeurs est donnée par

$$F(x) = \theta x + (1-\theta)F_1(x), \quad x \in [0,1],$$

et on note les densités correspondantes $f = \theta \mathbf{1}_{[0,1]} + (1-\theta)f_1$, où f_1 est une densité inconnue sur $[0,1]$. Lorsque la transformation T_i correspond à la z-valeur Z_i on obtient,

$$\begin{aligned} G_0(x) &= \mathbb{P}_{H_0}(Z_i \leq x) = \mathbb{P}_{H_0}(P_i \leq \Phi(x)) = \Phi(x), \\ \text{et } G_1(x) &= \mathbb{P}_{H_1}(Z_i \leq x) = \mathbb{P}_{H_1}(P_i \leq \Phi(x)) = F_1(\Phi(x)). \end{aligned}$$

Enfin dans ce cas, en termes de densité, on a $g(x) = \phi(x)[\theta + (1-\theta)f_1(\Phi(x))]$ pour tout $x \in \mathbb{R}$, où ϕ est la densité de la loi normale centrée et réduite.

Dans la suite, nous allons introduire différentes mesures de l'erreur des procédures de test multiple. Ces mesures sont naturellement reliées aux paramètres θ et f_1 ou g_1 que nous venons d'introduire.

1.3 Mesures de l'erreur

Une procédure de test multiple ou MTP fournit des régions de rejet, i.e. un ensemble de valeurs de la variable T_i pour lesquelles on décide de rejeter l'hypothèse nulle $H_i = 0$. En pratique, on obtient un ensemble R aléatoire, sous-ensemble des indices $\{1, \ldots, n\}$, tel que les indices sélectionnés correspondent aux hypothèses nulles qui sont rejetées. La procédure peut être définie à partir de la suite de p-valeurs $p = \{p_i, 1 \leq i \leq n\} \in [0,1]^n$ auquel cas la procédure de test multiple correspond à une fonction

$$R : p = (p_i)_{1 \leq i \leq n} \in [0,1]^n \mapsto R(p) \subset \{1, \ldots, n\}.$$

Un cas particulier est obtenu lorsque l'on considère les procédures dites de *seuillage*, de la forme $R(p) = \{i \in \{1, \ldots, n\}; p_i \leq t(p)\}$ où le seuil $t(\cdot) \in [0,1]$ peut dépendre des observations.

Un grand nombre de mesures d'erreur ont été proposées dans la litérature afin de mesurer la qualité d'une procédure de test multiple. Toutes cherchent à évaluer l'importance des hypothèses nulles rejetées à tort, ou taux de faux positifs (FP). Deux erreurs en particulier sont très utilisées : le *family wise error rate* ou FWER et le taux de fausses découvertes ou FDR (ce dernier étant parfois remplacé par la proportion de faux positifs, ou FDP). Nous rappelons ces différentes définitions en nous appuyant sur la table 1.

Le FWER est défini comme la probabilité de faire au moins un faux positif parmi toutes les hypothèses, soit

$$\text{FWER} = \mathbb{P}(\text{FP} \geq 1).$$

La proportion de fausses découvertes FDP est définie comme la proportion de faux positifs parmi les hypothèses rejetées,

$$\mathrm{FDP} = \frac{\mathrm{FP}}{\max(\mathrm{R}, 1)}.$$

On peut remarquer que le FDP est une variable aléatoire et ne défini pas un taux d'erreur. Dans [1], le taux de fausses découvertes ou FDR est défini comme l'espérance du FDP, i.e.

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{FP}}{\max(\mathrm{R}, 1)}\right] = \mathbb{E}\left[\frac{\mathrm{FP}}{\mathrm{R}}\Big|\mathrm{R} > 0\right]\mathbb{P}(\mathrm{R} > 0).$$

Le FDR est un critère de contrôle de test multiple beaucoup moins stringent que le FWER qui donne donc des procédures plus puissantes. Dans [15], Storey propose de modifier le FDR pour obtenir un nouveau critère, dénommé *positive FDR*, ou pFDR défini comme suit

$$\mathrm{pFDR} = \mathbb{E}\left[\frac{\mathrm{FP}}{\mathrm{R}}\Big|\mathrm{R} > 0\right].$$

Ce critère fait plus de sens que le précédent puisque lorsque l'on contrôle le FDR au niveau α et que des hypothèses ont été rejetées alors le FDR n'a été en réalité contrôlé qu'au niveau $\alpha\mathbb{P}(R > 0)$. Le pFDR ne présente pas ce défaut.

Dans la thèse, nous rappelons les différentes procédures de tests multiples qui ont été proposées dans la littérature et pour lesquelles des contrôles de taux d'erreur ont pu être établis. Un point important doit être noté : la plupart de ces MTP sont en fait des procédures de seuillage appliquées à des estimateurs des quantités FDR ou pFDR. La construction de procédures MTP dont on sait contrôler des taux d'erreur est donc intimement liée à l'estimation des quantités de type FDR ou pFDR. Or, ces quantités sont naturellement reliées à la proportion d'hypothèses nulles vraies θ et à la distribution sous l'alternative G_1 des statistiques de test X_i (resp. la distribution sous l'alternative F_1 de leur transformées T_i). Ainsi, comme déjà constaté plus haut, une MTP qui porte sur les p-valeurs a une région de rejet de la forme $[0, \gamma]$ pour un certain $\gamma > 0$. Alors, le pFDR de cette procédure s'écrit

$$pFDR(\gamma) = \frac{\theta\mathbb{P}(P \leq \gamma | H = 0)}{\mathbb{P}(P \leq \gamma)} = \frac{\theta\gamma}{F(\gamma)}.$$

Dans [4], les auteurs introduisent le concept de FDR local ou ℓFDR qui quantifie la plausibilité qu'une hypothèse soit vraie, sachant la valeur de sa statistique de test (ou de sa p-valeur). Dans le cadre du modèle de mélange, le ℓFDR peut être vu comme une probabilité a posteriori

$$\ell\mathrm{FDR}(x) = \mathbb{P}(H_i = 0 | X = x) = 1 - \frac{(1-\theta)g_1(x)}{\theta g_0(x) + (1-\theta)g_1(x)}.$$

Là encore, la mise au point d'une MTP par seuillage dont on contrôle l'erreur ℓFDR passe par l'estimation de cette quantité, donc par l'estimation des paramètres θ (proportion d'hypothèses nulles vraies) et g_1 ou f_1 (densités des variables X_i ou P_i sous l'alternative). Si on se replace dans le contexte des modèles de mélange exposé à la section précédente, on peut se poser les questions d'estimation semi-paramétrique par exemple des quantités (θ, f_1) qui définissent ces modèles. Ces estimateurs, une fois construit (et leurs propriétés étudiées), permettront par des méthodes dites de substitution de revenir au problème de la construction de procédures de tests multiples.

1.4 Inférence dans des modèles semi-paramétrique

Dans cette section, nous abordons brièvement la problématique de l'efficacité dans les modèles semi-paramétriques. Les notations sont celles du chapitre 25 de [19]. Les modèles semi-paramétriques sont de la forme $(\theta, f) \mapsto \mathbb{P}_{\theta,f}$, où θ est un paramètre euclidien tandis que f est non-paramétrique, i.e. appartient à un espace de dimension infinie (une classe de distributions par exemple). Nous nous intéressons ici au cas où l'on souhaite inférer la fonctionnelle $\psi(\mathbb{P}_{\theta,f}) = \theta$, i.e. la partie euclidienne du paramètre, considérant ainsi f comme un paramètre de nuisance. La théorie de l'efficacité asymptotique dans ces modèles est une extension naturelle de celle développée dans les modèles paramétriques.

On commence par rappeler la définition d'une fonction de score dans le cadre le plus général (non nécessairement semi-paramétrique). Soit $X_1, \ldots, X_n$ un échantillon de loi $\mathbb{P} \in \mathcal{P}$ à valeurs dans un ensemble $\mathcal{X}$.

Définition 1. *Un chemin différentiable est une application $t \mapsto \mathbb{P}_t$ définie d'un voisinage $[0, \epsilon)$ de 0 dans $\mathcal{P}$ avec $\mathbb{P}_0 = \mathbb{P}$, telle qu'il existe une fonction mesurable $g : \mathcal{X} \to \mathbb{R}$ satisfaisant*

$$\int \left(\frac{d\mathbb{P}_t^{1/2} - d\mathbb{P}^{1/2}}{t} - \frac{1}{2} g d\mathbb{P}^{1/2} \right) \to 0 \quad \textit{lorsque } t \to 0. \tag{2}$$

Le sous-modèle paramétrique $\{\mathbb{P}_t, 0 \leq t < \epsilon\}$ est alors dit différentiable en moyenne quadratique en $\mathbb{P}$ et g est la fonction de score du sous-modèle $\{\mathbb{P}_t, 0 \leq t < \epsilon\}$.

Lorsque $t \mapsto \mathbb{P}_t$ décrit une collection de tels sous-modèles, on obtient un ensemble de fonctions de score qui forment un espace tangent pour $\mathcal{P}$ en $\mathbb{P}_0$, noté $\dot{\mathcal{P}}_{\mathbb{P}}$. En considérant tous les chemins différentiables possibles on obtient une collection maximale de fonctions de score, appelée espace tangent maximal. On peut remarquer que très souvent, les chemins $t \mapsto \mathbb{P}_t$ sont construits tels que pour tout x, on ait

$$g(x) = \frac{\partial}{\partial t}\Big|_{t=0} \log d\mathbb{P}_t(x).$$

Cependant, cette propriété de différentiabilité ponctuelle n'est pas requise par la différentiabilité quadratique. Réciproquement, la différentiabilité ponctuelle n'assure pas (2). Plus de détails sur ce point sont donnés dans [19].

Afin de définir l'information disponible pour estimer $\psi(\mathbb{P})$, nous allons considérer uniquement les sous-modèles $t \mapsto \mathbb{P}_t$ le long desquels l'application $t \mapsto \psi(\mathbb{P}_t)$ est différentiable en $t = 0$ en un certain sens. Plus précisément, une application $\psi : \mathcal{P} \to \mathbb{R}$ est dite différentiable en $\mathbb{P}$ relativement à un espace tangent $\dot{\mathcal{P}}_{\mathbb{P}}$ s'il existe une application linéaire $\dot{\psi}_{\mathbb{P}} : \mathbb{L}_2(\mathbb{P}) \to \mathbb{R}$ telle que pour toute fonction de score $g \in \dot{\mathcal{P}}_{\mathbb{P}}$ et tout sous-modèle $t \mapsto \mathbb{P}_t$ associé à cette fonction de score, on ait

$$\frac{\partial \psi(\mathbb{P}_t)}{\partial t}\Big|_{t=0} = \lim_{t \to 0} \frac{\psi(\mathbb{P}_t) - \psi(\mathbb{P})}{t} = \dot{\psi}_{\mathbb{P}} g.$$

D'après le théorème de représentation de Riesz pour les espaces de Hilbert, il existe une fonction mesurable $\tilde{\psi}_{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ telle que

$$\dot{\psi}_{\mathbb{P}} g = < \tilde{\psi}_{\mathbb{P}}, g >_{\mathbb{L}_2(\mathbb{P})} = \int \tilde{\psi}_{\mathbb{P}} g d\mathbb{P}.$$

Une telle fonction est appelée fonction d'influence mais n'est pas entièrement définie par ψ et le modèle $\mathcal{P}$ de façon unique. C'est pourquoi on considèrera dans la suite l'unique fonction $\tilde{\psi}_{\mathbb{P}}$ contenue dans $\overline{\text{lin}}\dot{\mathcal{P}}_{\mathbb{P}}$ (la fermeture de l'espace linéaire engendré par l'espace tangent $\dot{\mathcal{P}}_{\mathbb{P}}$) et qui prolonge la précédente, appelée fonction d'influence efficace.

À partir de cette fonction d'influence efficace, on peut montrer une borne de Cramer-Rao et établir que $\mathbb{P}\tilde{\psi}^2_{\mathbb{P}}$ est la plus petite variance asymptotique qu'un estimateur de $\psi(\mathbb{P})$ puisse atteindre. Plus précisément, pour toute fonction de score g dans un ensemble tangent $\dot{\mathcal{P}}_{\mathbb{P}}$, on note $\mathbb{P}_{t,g}$ un sous-modèle de la forme $t \to \mathbb{P}_t$ qui a pour fonction de score g et le long duquel ψ est différentiable. Nous énonçons à présent le théorème LAM (local asymptotic minimax) donnant une borne inférieure du risque quadratique asymptotique de n'importe quel estimateur T_n de $\psi(\mathbb{P})$, voir Théorème 25.21 dans [19].

Théorème 1 (LAM, Local Asymptotic Minimax)**.** *Supposons que $\psi : \mathcal{P} \to \mathbb{R}$ est différentiable en $\mathbb{P}$ relativement à l'espace tangent $\dot{\mathcal{P}}_{\mathbb{P}}$ avec fonction d'influence efficace $\tilde{\psi}_{\mathbb{P}}$. Si l'espace $\dot{\mathcal{P}}_{\mathbb{P}}$ est un cône convexe alors pour toute suite d'estimateurs T_n de $\psi(\mathbb{P})$, on a*

$$\sup_I \liminf_{n\to+\infty} \sup_{g\in I} \mathbb{P}_{1/\sqrt{n},g}[\sqrt{n}(T_n - \psi(\mathbb{P}_{1/\sqrt{n},g}))]^2 \geq \mathbb{P}\tilde{\psi}^2_{\mathbb{P}},$$

où le premier supremum porte sur tous les sous-ensembles I finis de l'espace tangent $\dot{\mathcal{P}}_{\mathbb{P}}$.

D'autres propriétés peuvent être établies, comme le théorème de convolution. On renvoie à [19] pour plus de détails.

Avant de finir cette section, nous allons expliciter la forme des objets introduits ci-dessus dans le cadre d'un modèle semi-paramétrique où l'on s'intéresse de plus à la fonctionnelle $\psi(\mathbb{P}_{\theta,f}) = \theta$, i.e. la partie euclidienne de la paramétrisation. On considère donc $\mathcal{P} = \{\mathbb{P}_{\theta,f} : \theta \in \Theta, f \in \mathcal{F}\}$ où $\Theta \subset \mathbb{R}$ est un ouvert et $\mathcal{F}$ est un espace de densités (par rapport à la mesure de Lebesgue) de dimension infinie. Les sous-modèles considérés ici sont de la forme $t \to \mathbb{P}_{\theta+ta,f_t}$ pour tout chemin $t \to f_t$ dans $\mathcal{F}$ avec $f_0 = f$ et $a \in \mathbb{R}$. Les fonctions de score associées à ces chemins prennent typiquement une forme additive avec un terme de dérivée pour chaque composante θ ou f. Ainsi, si $\dot{l}_{\theta,f}$ est la fonction de score ordinaire du modèle pour θ lorsque f est fixée, on attend

$$\frac{\partial}{\partial t}\Big|_{t=0} \log d\mathbb{P}_{\theta+ta,f_t} = a\dot{l}_{\theta,f} + g.$$

La fonction g s'interprète comme un score pour f lorsque θ est fixé et varie typiquement dans un espace de dimension infinie. Cet espace est l'espace tangent pour f et on le note à présent ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$. La fonctionnelle $\psi(\mathbb{P}_{\theta,f}) = \theta + ta$ est différentiable (au sens ordinaire) par rapport à t et sa dérivée vaut a. Mais il faut davantage pour avoir la différentiabilité en $\mathbb{P}_{\theta,f}$ par rapport à l'espace tangent global $\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$. Par définition, ψ est différentiable relativement à $\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ si et seulement si il existe une fonction $\tilde{\psi}_{\theta,f}$ telle que

$$a = \frac{\partial}{\partial t}\Big|_{t=0} \psi(\mathbb{P}_{\theta+ta,f_t}) = < \tilde{\psi}_{\theta,f}, a\dot{l}_{\theta,f} + g >_{\mathbb{P}_{\theta,f}} \quad \forall a \in \mathbb{R}, \forall g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}.$$

En choisissant $a = 0$ ci-dessus, on constate que $\tilde{\psi}_{\theta,f}$ est nécessairement orthogonale à l'espace tangent pour le paramètre de nuisance ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$. On défini alors l'opérateur de projection

orthogonale $\Pi_{\theta,f} : \mathbb{L}_2(\mathbb{P}_{\theta,f}) \to \overline{\text{lin}}({}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}})$ sur (la fermeture de) l'espace linéaire engendré par cet espace tangent, puis la fonction

$$\tilde{l}_{\theta,f} = \dot{l}_{\theta,f} - \Pi_{\theta,f}\dot{l}_{\theta,f}$$

appelée fonction de score efficace pour θ et enfin sa covariance $\tilde{I}_{\theta,f} = \mathbb{P}_{\theta,f}\tilde{l}^2_{\theta,f}$ appelée matrice d'information efficace pour θ. On obtient alors le lemme suivant.

Lemme 1. *Supposons que pour tout $a \in \mathbb{R}$ et tout $g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ il existe un chemin $t \to f_t$ dans $\mathcal{F}$ tel que*

$$\int \left[\frac{d\mathbb{P}^{1/2}_{\theta+ta,f_t} - d\mathbb{P}^{1/2}_{\theta,f}}{t} - \frac{1}{2}(a\dot{l}_{\theta,f} + g)d\mathbb{P}^{1/2}_{\theta,f} \right]^2 \to 0 \text{ lorsque } t \to 0.$$

Si de plus $\tilde{I}_{\theta,f}$ est inversible, alors la fonction $\psi(\mathbb{P}_{\theta,f}) = \theta$ est différentiable en $\mathbb{P}_{\theta,f}$ relativement à l'espace tangent $\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}} = lin\ \dot{l}_{\theta,f} + {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}} = \{a\dot{l}_{\theta,f} + g : a \in \mathbb{R}, g \in {}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}\}$ avec fonction d'influence efficace $\tilde{\psi}_{\theta,f} = \tilde{I}^{-1}_{\theta,f}\tilde{l}_{\theta,f}$.

Une conséquence de ce résultat est que lorsque de plus l'espace tangent pour le paramètre de nuisance ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$ est un cône convexe, la variance asymptotique optimale de toute suite d'estimateurs T_n de θ est $\tilde{I}^{-1}_{\theta,f}$. Ainsi, en présence du paramètre de nuisance f, une partie de l'information disponible pour estimer θ peut être perdue, et cette partie est quantifiée par la projection orthogonale de la fonction de score $\dot{l}_{\theta,f}$ pour θ sur l'espace tangent pour le paramètre de nuisance ${}_f\dot{\mathcal{P}}_{\mathbb{P}_{\theta,f}}$. Plus de détails sont donnés dans [19] et dans la thèse.

1.5 Organisation du manuscrit

Dans toute la thèse, on suppose que les statistiques de test $X_1, \ldots, X_n$ sont i.i.d. de distribution continue sous l'hypothèse nulle comme sous l'alternative. Alors les p-valeurs associées $P_1, \ldots, P_n$ sont également i.i.d. et sous l'hypothèse nulle, suivent une distribution uniforme sur l'intervalle $[0, 1]$ notée $\mathcal{U}[0, 1]$. Alors, la densité g de ces p-valeurs (définie par rapport à la mesure de Lebesgue sur $[0, 1]$) est donnée par un modèle de mélange avec deux composantes,

$$\forall x \in [0, 1] \quad g(x) = \theta + (1 - \theta)f(x), \tag{3}$$

où $\theta \in [0, 1]$ est inconnu et représente la proportion d'hypothèses nulles qui sont vraies dans l'échantillon, alors que f représente la densité - également inconnue- des p-valeurs sous l'hypothèse alternative. Dans ce contexte, on cherche à estimer les paramètres (θ, f) dans le but d'utiliser ces estimateurs afin de construire des procédures MTP optimales. C'est l'objet des chapitres 2 et 3 de la thèse dont les résultats sont présentés plus en détails ci-dessous.

2 Première contribution : estimation de la proportion de vraies hypothèses nulles

Dans le chapitre 2 de la thèse, on se concentre sur l'estimation du paramètre euclidien θ dans le modèle (3) ci-dessus. Un grand nombre d'estimateurs de cette proportion θ ont

été proposés dans la littérature mais il y a relativement peu d'études théoriques sur les propriétés de convergence de ces estimateurs.

Commençons tout d'abord par nous intéresser à la question de l'identifiabilité des paramètres dans le modèle (3). Nous établissons la propriété suivante.

Proposition 1. *Le paramètre (θ, f) est identifiable sur un ensemble $(0,1) \times \mathcal{F}$ si et seulement si pour tout $f \in \mathcal{F}$ et tout $c \in (0,1)$ on a $c + (1-c)f \in \mathcal{F}$.*

La proposition précédente possède l'avantage d'établir une condition à la fois nécessaire et suffisante pour l'identifiabilité des paramètres. Bien qu'elle ne soit pas explicite, elle permet un éclairage nouveau sur les différentes hypothèses (suffisantes mais pas nécessaires) faites dans la littérature et impliquant l'identifiabilité des paramètres. Ainsi dans [5], les auteurs introduisent la condition de *pureté* qui impose

$$\inf_{x \in [0,1]} f(x) = 0.$$

On peut ainsi voir que si l'espace $\mathcal{F}$ ne contient que des densités qui satisfont à l'hypothèse de pureté, alors la condition d'identifiabilité de la Proposition 1 est satisfaite. De même, dans [6], les auteurs imposent que f est décroissante, avec $f(1) = 0$ et dans [2, 10] la densité f s'annule sur un voisinage de 1 ou sur un intervalle de $(0,1)$. Dans tous ces cas, on voit que les hypothèses faites sont suffisantes pour assurer l'identifiabilité.

Dans la suite de ce travail, on s'intéresse au cas où la densité sous l'alternative appartient à l'ensemble suivant

$$\mathcal{F}_\lambda = \{f : [0,1] \to \mathbb{R}^+, \text{ densité continue et décroissante, positive sur } [0,\lambda) \text{ avec } f_{|[\lambda,1]} = 0\},$$

pour un certain $\lambda \in (0,1]$. Dans le cas où $\lambda = 1$, on retrouve les hypothèses de pureté ou celle introduite dans [6]. On notera que l'hypothèse de décroissance n'est pas nécessaire à la suite des développements mais les résultats obtenus sont plus forts avec cette hypothèse supplémentaire.

La littérature sur l'estimation du paramètre θ peut grossièrement se diviser en 3 types de méthodes : les estimateurs basés sur des histogrammes ; ceux basés sur des estimateurs de densité monotones ; et enfin ceux basés sur des estimateurs de densité régulières. On constate que pour le premier type d'estimateur (par histogrammes), très peu de résultats de convergence ont été établis. La convergence en probabilité d'un estimateur de ce type est prouvée dans [2]. Quant à l'estimateur proposé dans [13] et défini de la façon suivante

$$\hat{\theta}_n(\lambda) = \frac{\sharp\{P_i > \lambda, 1 \le i \le n\}}{n(1-\lambda)},$$

où λ est ici un paramètre à choisir, on peut constater qu'une version oracle de cet estimateur qui utiliserait la vraie valeur de λ a des propriétés de normalité asymptotique. Plus précisément, si on suppose que $f_{|[\lambda^\star,1]} = 0$ et que l'on choisi $\lambda = \lambda^\star$, alors on obtient

$$\sqrt{n}[\hat{\theta}_n(\lambda^\star) - \theta] \to^d_{n \to +\infty} \mathcal{N}\Big(0, \theta\Big(\frac{1}{1-\lambda^\star} - \theta\Big)\Big).$$

Les autres types d'estimateurs (par densité régulière ou monotone), convergent quant à eux à des vitesses non paramétriques [6, 9].

Les questions que nous nous sommes posées sont les suivantes : quand est-il possible de construire un estimateur de θ qui converge à vitesse paramétrique ? Quelle est la variance asymptotique optimale d'un estimateur paramétrique ? Existe-t-il des estimateurs efficaces ?

Les résultats que nous avons établis sont les suivants. Nous rappelons que nous supposons $f \in \mathcal{F}_\lambda$. Deux cas sont à distinguer ici, suivant que $\lambda = 1$ ou $\lambda < 1$. Lorsque $\lambda = 1$, nous prouvons qu'aucun estimateur de θ ne peut atteindre la vitesse de convergence paramétrique. Ainsi, lorsque f ne s'annule pas en plus qu'un point, seules les vitesses non-paramétriques peuvent être atteintes. Dans le cas où $\lambda < 1$, nous montrons qu'il est possible de construire des estimateurs convergents à vitesse paramétrique, nous en exhibons, mais ces estimateurs n'atteignent pas la variance asymptotique optimale. Nous conjecturons de plus que dans les modèles « réguliers », il n'est pas possible d'atteindre cette variance optimale et qu'aucun estimateur n'est efficace.

3 Seconde contribution : estimation de la densité sous l'alternative

Dans cette partie, on suppose que l'on dispose d'un estimateur de la proportion θ et on s'intéresse à l'estimation de la densité f. Plusieurs modélisations pour f ont été considérées dans la littérature, aussi bien paramétriques que non-paramétriques. Nous nous concentrons ici sur le cadre non-paramétrique qui a l'avantage d'être très général. Dans [6], les auteurs ont proposé un estimateur de Grenander pour densités monotones pour estimer f, modifié ensuite dans [17]. Une autre approche a consisté en des hypothèses de régularité sur f et la construction d'estimateurs à noyau. C'est le cas par exemple dans [9] où le but original est d'estimer θ et non pas f. Un autre estimateur à noyau a été proposé dans [11] ainsi qu'une procédure de test multiple, appelée `kerfdr`. Il s'agit d'un algorithme itératif inspiré de l'algorithme `em` dont les auteurs établissent la convergence lorsque le nombre d'itérations tend vers l'infini. Cependant, la procédure proposée n'optimise aucun critère (contrairement à l'algorithme `em`) et en particulier, elle ne permet pas a priori d'augmenter la vraisemblance. De plus, les propriétés asymptotiques (avec la taille de l'échantillon) de l'estimateur proposé ne sont pas étudiées. Ceci est dû en partie au fait que la forme itérative de l'algorithme rend difficile l'analyse de l'estimateur produit.

Nous avons dans une première partie de ce travail proposé un estimateur à noyau avec poids aléatoires qui est très proche (par l'esprit au moins) de l'estimateur proposé [11]. Le but est ici de valider en quelque sorte l'approche `kerfdr` et de donner un éclairage sur les propriétés de convergence (lorsque la taille de l'échantillon augmente) de cette procédure.

Supposons que l'on dispose d'un estimateur préliminaire $\hat{\theta}_n$ de θ ainsi que d'un estimateur non-paramétrique $\hat{g}_n$ de la densité g des observations dans le modèle (3). Nous proposons ici de considérer un estimateur à noyau de g défini de la façon suivante

$$\hat{g}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) = \frac{1}{n}\sum_{i=1}^{n} K_{i,h}(x), \tag{4}$$

où K est un noyau et $h > 0$ une fenêtre (qui seront choisis plus tard) avec

$$K_{i,h}(\cdot) = \frac{1}{h} K\Big(\frac{\cdot - X_i}{h}\Big). \tag{5}$$

On peut remarquer que cet estimateur de g est consistant sous des hypothèses raisonnables.

À partir de $(\hat{\theta}_n, \hat{g}_n)$, plusieurs approches sont possibles pour estimer la densité f. L'approche naïve consiste à définir

$$\hat{f}_n^{\text{naive}}(x) = \frac{\hat{g}_n(x) - \hat{\theta}_n}{1 - \hat{\theta}_n} \mathbf{1}_{\{\hat{\theta}_n \neq 1\}},$$

où 1_A est la fonction indicatrice de l'ensemble A. Nous montrons que cet estimateur a les mêmes propriétés théoriques que l'estimateur à noyau avec poids aléatoires défini ci-dessous, cependant ses performances sur données simulées sont bien plus mauvaises. Une autre construction moins naïve pour estimer f est la suivante. Pour chaque hypothèse testée H_i, on défini une variable aléatoire latente Z_i qui prend les valeurs suivantes

$$\forall i = 1, \dots, n \quad Z_i = \begin{cases} 0 & \text{si } H_i \text{ est vraie,} \\ 1 & \text{sinon.} \end{cases}$$

De façon intuitive, il serait intéréssant d'utiliser la valeur de Z_i comme poids pour pondérer les observations X_i à utiliser dans la construction de l'estimateur à noyau. Cette idée conduirait à la définition suivante

$$f_1(x) = \frac{1}{h} \sum_{i=1}^{n} \frac{Z_i}{\sum_{k=1}^{n} Z_k} K\Big(\frac{x - X_i}{h}\Big) = \sum_{i=1}^{n} \frac{Z_i}{\sum_{k=1}^{n} Z_k} K_{i,h}(x), \ x \in [0,1].$$

Cependant, f_1 n'est pas un estimateur puisque les variables Z_i ne sont pas observées. L'approche naturelle (initialement proposée dans [11]) est de remplacer les Z_i par leur espérance conditionnelle aux données. Ainsi, on introduit $\tau(X_i) = \mathbb{E}(Z_i | X_i)$ défini par

$$\forall x \in [0,1], \ \tau(x) = \mathbb{E}(Z_i | X_i = x) = \frac{(1-\theta) f(x)}{g(x)} = 1 - \frac{\theta}{g(x)}, \tag{6}$$

ce qui permet de définir

$$\forall x \in [0,1], \ f_2(x) = \sum_{i=1}^{n} \frac{\tau(X_i)}{\sum_{k=1}^{n} \tau(X_k)} K_{i,h}(x). \tag{7}$$

Là encore, f_2 n'est pas un estimateur puisque les espérances conditionnelles τ utilisées en tant que poids dépendent des paramètres du modèle qui sont inconnus. Afin de résoudre ce problème, les auteurs de [11] ont proposé une approche itérative, appelée `kerfdr`, afin d'approximer (7). Nous proposons quant à nous de remplacer ces poids par des estimateurs. Ainsi, soit

$$\forall x \in [0,1], \ \hat{\tau}(x) = 1 - \frac{\hat{\theta}_n}{\hat{g}_n(x)}. \tag{8}$$

Alors on peut définir les poids de la façon suivante

$$\hat{\tau}_i = \hat{\tau}(X_i) = 1 - \frac{\hat{\theta}_n}{\tilde{g}_n(X_i)}, \ \text{où} \ \ \tilde{g}_n(X_i) = \frac{1}{(n-1)} \sum_{j \neq i}^{n} K_{j,h}(X_i), \tag{9}$$

et obtenir un estimateur à noyau de f avec poids aléatoires

$$\forall x \in [0,1],\ \hat{f}_n^{\mathrm{rwk}}(x) = \sum_{i=1}^{n} \frac{\hat{\tau}_i}{\sum_{k=1}^{n} \hat{\tau}_k} K_{i,h}(x). \tag{10}$$

Les liens entre cet estimateur et l'estimateur itératif `kerfdr` sont discutés plus en détails dans le chapitre 3 de la thèse. Nous étudions la convergence de cet estimateur sur l'ensemble des densités Hölderiennes.

Définition 2. *Soit $\beta > 0, L > 0$ et $H(\beta, L)$ l'ensemble des fonctions $\psi : [0,1] \to \mathbb{R}$ qui sont l-fois continûment différentiables sur $[0,1]$ avec $l = \lfloor \beta \rfloor$ et telles que*

$$|\psi^{(l)}(x) - \psi^{(l)}(y)| \leq L|x-y|^{\beta-l}, \quad \forall x, y \in [0,1].$$

L'ensemble $H(\beta, L)$ est appelé (β, L)-classe de Hölder.

On note ensuite $\Sigma(\beta, L)$ l'ensemble

$$\Sigma(\beta, L) = \Big\{ \psi : \psi \text{ densité sur } [0,1] \text{ et } \psi \in H(\beta, L) \Big\}.$$

On établit alors le théorème suivant.

Théorème 2. *Pour un bon choix du noyau K, si $\hat{\theta}_n$ converge presque sûrement vers θ et pour la fenêtre $h = \alpha n^{-1/(2\beta+1)}$ avec $\alpha > 0$, on obtient que pour tout $\delta > 0$, le risque quadratique ponctuel $\hat{f}_n^{rwk}$ satisfait*

$$\sup_{x\in[0,1]} \sup_{\theta\in[\delta,1-\delta]} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_{\theta,f}(|\hat{f}_n^{rwk}(x) - f(x)|^2) \leq C_1 \sup_{\theta\in[\delta,1-\delta]} \sup_{f\in\Sigma(\beta,L)} \left[\mathbb{E}_{\theta,f}\left(|\hat{\theta}_n - \theta|\right)^4\right]^{\frac{1}{2}} + C_2 n^{\frac{-2\beta}{2\beta+1}},$$

où C_1, C_2 sont deux constantes positives qui dépendent uniquement de β, L, α, δ et K.

Dans une seconde partie de ce travail, on étudie un estimateur de maximum de vraisemblance régularisé. En suivant le travail de [7], on construit en fait une suite itérative d'estimateurs de la densité f à partir de la maximisation d'une vraisemblance régularisée. Dans la suite, on suppose que K est un noyau positif et symétrique sur $\mathbb{R}$ et on définit sa version normalisée

$$K_h(x) = h^{-1} K(h^{-1}x).$$

On considère également l'opérateur linéaire de régularisation $\mathcal{S} : \mathbb{L}_1([0,1]) \to \mathbb{L}_1([0,1])$ défini par

$$\mathcal{S}f(x) = \int_0^1 \frac{K_h(u-x)f(u)}{\int_0^1 K_h(s-u)ds} du, \text{ pout tout } x \in [0,1].$$

On remarque que si f est une densité sur $[0,1]$ alors c'est le cas aussi de $\mathcal{S}f$. On va à présent faire l'hypothèse supplémentaire sur la densité f suivante : $f \in \mathcal{F}$ avec

$$\mathcal{F} = \{\text{densités } f \text{ sur } [0,1] \text{ t.q. } \log f \in \mathbb{L}_1([0,1])\}.$$

On défini ensuite l'opérateur adjoint $\mathcal{S}^* : \mathbb{L}_1([0,1]) \to \mathbb{L}_1([0,1])$ via

$$\mathcal{S}^* f(x) = \frac{\int_0^1 K_h(u-x) f(u) du}{\int_0^1 K_h(s-x) ds}.$$

Toute densité $f \in \mathcal{F}$ est alors approchée par un opérateur de régularisation non-linéaire $\mathcal{N}$ défini par

$$\mathcal{N} f(x) = \exp\{(\mathcal{S}^*(\log f))(x)\}, \quad x \in [0,1].$$

On peut remarquer que $\mathcal{N} f$ n'est pas nécessairement une densité. Notre procédure de maximum de vraisemblance régularisée consiste alors en une procédure itérative qui utilise un estimateur à noyau avec poids aléatoires dont les poids sont itérativement remis à jour, et définis à partir de l'estimateur régularisé $\mathcal{N}\hat{f}$. Plus de détails sont fournis sur la construction exacte de cette procédure dans le chapitre 3 du manuscrit.

Le résultat principal concernant cette procédure est une propriété de descente de l'algorithme qui assure sa convergence (avec le nombre d'itérations). Cette procédure ainsi que les précédentes sont alors utilisées dans le contexte de l'estimation du ℓFDR et nous étudions leurs performances sur des simulations.

Références

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) :289–300, 1995.

[2] Alain Celisse and Stéphane Robin. A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference*, 140(11) :3132–3147, 2010.

[3] Sandrine Dudoit and Mark J. van der Laan. *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. Springer, New York, 2008.

[4] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456) :1151–1160, 2001.

[5] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3) :1035–1061, 2004.

[6] Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(4) :555–572, 2005.

[7] M. Levine, D. R. Hunter, and D. Chauveau. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2) :403–416, 2011.

[8] Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1) :373–393, 2006.

[9] Pierre Neuvial. Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. *J. Machine Learning Research*, pages 1423–1459, 2013.

[10] Stan Pounds and Cheng Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16) :1979–1987, 2006.

[11] Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre. A semi-parametric approach for mixture models : application to local false discovery rate estimation. *Comput. Statist. Data Anal.*, 51(12) :5483–5493, 2007.

[12] Etienne Roquain. Type I error rate control for testing many hypotheses : a survey with proofs. *J. SFdS*, 152(2) :3–38, 2011.

[13] T. Schweder and E. Spjøtvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3) :493–502, 1982.

[14] John D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3) :479–498, 2002.

[15] John D. Storey. The positive false discovery rate : a Bayesian interpretation and the q-value. *Ann. Statist.*, 31(6) :2013–2035, 2003.

[16] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1) :187–205, 2004.

[17] Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1) :303, 2008.

[18] F.E. Turkheimer, C.B. Smith, and K. Schmidt. Estimation of the number of true null hypotheses in multivariate analysis of neuroimaging data. *NeuroImage*, 13(5) :920 – 930, 2001.

[19] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.