L'astuce bayésienne en test multiple

Etienne Roquain [1]

[1] *Sorbonne Universités, UPMC Univ Paris 6, UMR 7599, LPMA, F-75005, Paris, France*

Résumé. Cette note présente des procédures de Bayes empiriques pour le problème de contrôle du taux moyen de faux positifs (FDR). Il est montré notamment que les analogues bayésiens des p-values, appelés q-values (locales ou non), peuvent être utilisés sans correction de tests multiples, ce qui contraste fortement avec l'approche classique. Cette note est une synthèse basée sur les travaux de Efron et al. (2001), Storey (2003) et Sun et Cai (2009).

Mots-clés. Bayes empirique, taux de fausses découvertes, méthode de Benjamini-Hochberg, p-values, q-values.

Abstract. This short note presents empirical Bayes procedures for the problem of controlling the false discovery rate (FDR). We show that the Bayesian counterpart of the p-values, called the (local) q-values, can be used without multiple testing correction. This is markedly different from the classical p-value approach. This note is an unifying study based on the work of Efron et al. (2001), Storey (2003) and Sun et Cai (2009).

Keywords. Empirical Bayes, False discovery rate, Benjamini-Hochberg method, p-values, q-values.

1 Contexte

Pour le statisticien moderne, une tâche quotidienne consiste à extraire les éléments significatifs de données, possiblement de grandes dimensions, en effectuant m tests simultanément. Le nombre m de tests peut aller jusqu'à plusieurs millions, notamment pour traiter les données GWAS (Genome-wide association studies). Pour de telles valeurs de m, il est aisé de voir que rejeter toutes les hypothèses avec une p-value plus petite que α (disons 5%) peut faire "exploser" le nombre de faux positifs. Ainsi, *une correction de tests multiples s'impose.*

La correction la plus connue est certainement celle de Bonferroni, consistant à faire tous les tests au niveau α/m (au lieu de α). Elle est pourtant réputée trop conservative en pratique, dans le sens où elle ne permet pas de faire beaucoup de détections. Benjamini et Hochberg (1995) ont proposé une procédure, noté BH dans la suite, qui utilise un seuil intermédiaire entre la procédure non-corrigée (α) et celle de Bonferroni (α/m), précisément calibré pour contrôler la moyenne du taux de faux positifs (FDR) au niveau

α. En d'autres termes, dans l'ensemble des éléments déclarés significatifs, la proportion d'éléments erronés est (en moyenne) plus petite que α. La procédure BH est maintenant souvent considérée comme la correction de test multiple "par défaut" et elle est utilisée au sein d'une impressionnante variété de domaines d'application.

Parallèlement à cette approche, d'autres auteurs ont exploré des méthodes de tests multiples de nature bayésienne, voir Efron et al. (2001), Storey (2003) and Sun et Cai (2009). L'approche bayésienne empirique remonte à Robbins (1956) et peut être vue comme un moyen de choisir la loi a priori. Dans le cadre particulier du test multiple, elle permet de ne plus voir la multiplicité comme un "fléau" mais plutôt comme une opportunité pour ajuster une loi a priori.

2 Cadre mathématique

Test multiple

Commençons par décrire un cadre simple : supposons que l'on observe $X \in \mathbb{R}^m$, avec

$$X_i = \mu_i + \varepsilon_i, \quad 1 \leq i \leq m,$$

où ε_i , $1 \leq i \leq m$, sont i.i.d. $\mathcal{N}(0,1)$. On considère le problème du test des hypothèses nulles "$\mu_i = 0$" contre "$\mu_i \neq 0$", simultanément en $1 \leq i \leq m$. Pour simplifier, on peut supposer $\mu_i = H_i\Delta$ pour $\Delta \in \mathbb{R}\backslash\{0\}$ et $H = (H_i, 1 \leq i \leq m) \in \{0,1\}^m$, ce dernier étant le paramètre d'intérêt. Les statistiques de test usuelles sont $|X_i|$, $1 \leq i \leq m$ de p-values respectives définies par

$$p_i(X) = 2\overline{\Phi}(|X_i|), \quad 1 \leq i \leq m,$$

où $\overline{\Phi}(x) = \mathbb{P}(Z \geq x)$, $Z \sim \mathcal{N}(0,1)$. L'objectif est de retrouver H à partir de la famille de p-values[1].

On peut considérer le même objectif dans le cadre plus général suivant : on observe X (sur un espace mesurable) et on dispose d'une famille de p-values $(p_i(X), 1 \leq i \leq m)$, qui sont mutuellement indépendantes et de lois marginales liées à un vecteur $H \in \{0,1\}^m$ de la façon suivante

$$\begin{cases} p_i(X) \sim U(0,1) & \text{si } H_i = 0, \\ p_i(X) \sim F & \text{si } H_i = 1, \end{cases} \tag{1}$$

où F appartient à une certaine classe $\mathcal{F}$ de fonctions de répartition sur $[0,1]$. Par exemple, dans le cas particulier gaussien plus haut $\mathcal{F} = \{F : x \mapsto \overline{\Phi}(\overline{\Phi}^{-1}(x/2) - \Delta) + \overline{\Phi}(\overline{\Phi}^{-1}(x/2) + \Delta), \Delta \in \mathbb{R}\backslash\{0\}\}$.

[1] On renvoie le lecteur curieux à la page web interactive `http://simulations.lpma-paris.fr/fdr_tutorial/` pour quelques simulations.

Procédure et taux moyen de faux positifs (FDR)

Une procédure est définie simplement comme une application mesurable de la forme $\phi_i(X) \in \{0, 1\}$, $1 \leq i \leq m$. Des procédures particulières consistent à "faire du seuillage" :

$$\phi_i(X) = \mathbf{1}\{p_i(X) \leq t\}, \ 1 \leq i \leq m, \tag{2}$$

pour un seuil $t \in \mathbb{R}$ donné, possiblement aléatoire. Le risque (multiple) de type I de ces procédures peut être mesuré par les taux moyens de faux positifs suivants [2] :

$$\mathrm{FDR}(\phi) = \mathbb{E}\left[\frac{\sum_{i=1}^m (1 - H_i)\phi_i(X)}{\sum_{i=1}^m \phi_i(X)}\right], \quad \mathrm{pFDR}(\phi) = \frac{\mathrm{FDR}(\phi)}{\mathbb{P}(\sum_{i=1}^m \phi_i(X) > 0)}. \tag{3}$$

Notons que $\mathrm{FDR}(\phi)$ et $\mathrm{pFDR}(\phi)$ se distinguent par le fait que le pFDR ne fait la moyenne du taux de faux positifs que sur les événements où la procédure "a trouvé quelque chose".

Cadre bayésien

La difficulté mathématique pour manipuler le (p)FDR tient souvent au dénominateur situé à l'intérieur de l'espérance dans (3), qui dépend de l'observation X. Le cadre bayésien lève cette difficulté en se plaçant conditionnellement à X. En contrepartie, nous devons introduire une loi a priori sur les paramètres du modèle, à savoir (H, F).

La loi a priori sur H, notée ν_1, va définir la structure particulière des hypothèses nulles. La loi a priori sur F, notée ν_2, est d'une autre nature et nous supposerons simplement que $\nu_2 = \delta_F$ pour un $F \in \mathcal{F}$.

Par conséquent, dans cette note, les espérances seront désormais prises par rapport à la loi jointe de $(X, (H, F))$ avec une loi de X sachant H, F donnée par (1) et $(H, F) \sim \nu_1 \otimes \nu_2$.

3 Cas d'hypothèses nulles non structurées

Nous cherchons à obtenir une procédure qui possède un bon comportement lorsque les hypothèses n'ont aucune structure prédéfinie. Il paraît ainsi adapté de prendre $\nu_1 = \mathcal{B}(1 - \pi_0)^{\otimes m}$ comme loi a priori pour H, où $\pi_0 \in [0, 1]$ et $\mathcal{B}$ désigne la loi binomiale.

Procédure de Bayes

Lemme 3.1 *Parmi les procédures de la forme (2) avec un seuil $t \in \mathbb{R}$ déterministe qui assurent $pFDR(\phi) \leq \alpha$ sous l'a priori $\nu_1 = \mathcal{B}(1 - \pi_0)^{\otimes m}$ et $\nu_2 = \delta_F$, la procédure la plus puissante a pour seuil*

$$t_\alpha(\pi_0, F) = \max\{t \in [0, 1] \ : \ \pi_0 t / F(t) \leq \alpha\}. \tag{4}$$

[2] La convention $0/0 = 0$ est adoptée ici.

Pour prouver ce lemme, nous montrons tout d'abord que pour tout $t \in [0,1]$,

$$\mathrm{pFDR}(\phi) = \mathbb{P}(H_i = 0 \mid p_i(X) \le t) = \pi_0 t / F(t), \quad (5)$$

pour un i arbitraire. Notons $R(t) = \sum_{i=1}^m \mathbf{1}\{p_i(X) \le t\}$. Par le théorème de Fubini,

$$\mathrm{FDR}(\phi) = \mathbb{E}\left[\frac{1}{R(t)} \sum_{i=1}^m \mathbb{P}(H_i = 0 \mid \mathbf{1}\{p_j(X) \le t\}, 1 \le j \le m) \mathbf{1}\{p_i(X) \le t\} \mathbf{1}\{R(t) \ge 1\}\right]$$

D'après l'hypothèse d'indépendance, $\mathbb{P}(H_i = 0 \,|\, \mathbf{1}\{p_j(X) \le t\}, 1 \le j \le m) \mathbf{1}\{p_i(X) \le t\} = \mathbb{P}(H_i = 0 \mid \mathbf{1}\{p_i(X) \le t\}) \mathbf{1}\{p_i(X) \le t\} = \mathbb{P}(H_i = 0 \mid p_i(X) \le t) \mathbf{1}\{p_i(X) \le t\}$. Ceci conduit facilement à (5). La conclusion du lemme vient alors de ce que, pour tout seuil t déterministe assurant $\mathrm{pFDR}(\phi) \le \alpha$, nous avons $\pi_0 t / F(t) \le \alpha$. Notons au passage que le supremum dans (4) est bien un maximum. □

D'après le Lemme 3.1, la procédure de Bayes $t_\alpha(\pi_0, F)$ rejette notamment toutes les hypothèses satisfaisant $q_i(X) \le \alpha$, où

$$q_i(X) = \pi_0 p_i(X) / F(p_i(X)), \quad 1 \le i \le m. \quad (6)$$

Storey (2003) nomme la quantité (6) la *q-value*. Elle est souvent présentée comme l'analogue bayésien de la p-value. Le Lemme 3.1 indique donc que cette quantité ne nécessite plus de correction de test multiple !

Procédure de Bayes empirique

La procédure définie dans la section précédente dépend des hyper-paramètres π_0 et F. Une stratégie est de les estimer par des estimateurs $\widehat{\pi}_0$ et $\widehat{F}$ et de les "plugger" dans le seuil $t_\alpha(\pi_0, F)$ défini dans (4) : la procédure de Bayes empirique associée est donc définie par

$$t_\alpha(\widehat{\pi}_0, \widehat{F}) = \max\{t \in [0,1] \;:\; \widehat{F}(t) \ge \widehat{\pi}_0 t / \alpha\}, \quad (7)$$

où $\widehat{\pi}_0$ et $\widehat{F}$ sont deux estimateurs des hyper-paramètres π_0 et F, respectivement.

Lemme 3.2 *La procédure de seuil $t_\alpha(\widehat{\pi}_0, \widehat{F})$ définie par (7) et utilisant $\widehat{\pi}_0 = 1$ et $\widehat{F}(t) = m^{-1} \sum_{i=1}^m \mathbf{1}\{p_i(X) \le t\}$, $t \in [0,1]$, correspond exactement à la procédure BH.*

Ainsi, la procédure BH est une procédure de Bayes empirique. Comme l'ont montré Benjamini et Hochberg (1995), elle contrôle par ailleurs le FDR au niveau α pour tout H (sans loi a priori). Ainsi, il s'agit d'une procédure particulièrement adaptée lorsque les hypothèses nulles sont non structurées.

4 Cas d'hypothèses nulles structurées

La démarche bayésienne permet à la fois d'imposer une structure sur H via la loi a priori ν_1 et de donner une méthode de correction spécialement dédiée à cette structure.

Procédure de Bayes

La première étape est de changer les statistiques de tests. Une approche est de choisir celles qui sont optimales pour minimiser le risque de classification (pondéré) : $\mathcal{R}_\lambda(\phi) = \mathbb{E}\left(\lambda \sum_{i=1}^m (1-H_i)\phi_i + \sum_{i=1}^m H_i(1-\phi_i)\right)$, $\lambda > 0$. Classiquement, ce risque est minimum pour la règle de décision $\phi_i(X) = \mathbf{1}\{\ell_i(X) \leq (1+\lambda)^{-1}\}$ où

$$\ell_i(X) = \mathbb{P}(H_i = 0 \mid X), \quad 1 \leq i \leq m. \tag{8}$$

Les quantités $\ell_i(X)$ sont de nouvelles p-values de nature bayésienne, qui peuvent être interprétées comme des q-values “locales”. Elles peuvent être utilisées pour contrôler le FDR (toujours intégré par rapport aux loi a priori) de la façon suivante :

Lemme 4.1 *Ordonnons les q-values locales* (8) *dans l'ordre croissant* $\ell_{(1)}(X) \leq \cdots \leq \ell_{(m)}(X)$ *et posons*

$$\widehat{k} = \max\left\{k \in \{0, \ldots, m\} \;:\; k^{-1} \sum_{k'=1}^{k} \ell_{(k')}(X) \leq \alpha\right\}. \tag{9}$$

Alors la procédure $\phi_\alpha(\pi_0, F)$ *qui rejette les hypothèses correspondant aux* $\widehat{k}$ *plus petites q-values locales contrôle le FDR au niveau* α*, c'est-à-dire, FDR(*$\phi_\alpha(\pi_0, F)$*)* $\leq \alpha$ *pour toutes valeurs des hyper-paramètres* π_0 *et* F*.*

D'après le Lemme 4.1, Toutes les hypothèses satisfaisant $\ell_i(X) \leq \alpha$ sont rejetées par la procédure de Bayes $\phi_\alpha(\pi_0, F)$ du Lemme 4.1. Ainsi, de manière similaire aux q-values (6), les q-values locales (8) ne nécessitent pas de correction de test multiple. Notons que la procédure de Bayes rejette même un certain nombre d'hypothèses avec une q-value locale plus grande que α.

Calcul des q-values locales et procédures de Bayes empiriques

Pour pouvoir utiliser la procédure de Bayes $\phi_\alpha(\pi_0, F)$, il s'agit à présent de calculer les q-values locales et d'estimer les hyper-paramètres. Deux exemples sont donnés ci-dessous dans le cadre gaussien simple décrit en Section 2.

<u>Cas extrême</u> : Lorsque $H_i = H_j$ pour $i \neq j$ et $H_i \sim \mathcal{B}(1-\pi_0)$, toutes les statistiques de test $\ell_i(X)$ sont égales, disons à $\ell(X)$, donc $\phi_\alpha(\pi_0, F)_i = \mathbf{1}\{\ell(X) \leq \alpha\}$ pour tout i. Dans ce cas, il n'est pas difficile de voir que $\ell(X)$ est une bijection décroissante de $|\overline{X}_n|$ (après plug-in de $\widehat{\mu} = \overline{X}_n$), ce qui correspond à ce qui est attendu lors d'un test simple (bayésien). Notons que le seuil final dépend de π_0 qui ne peut pas être estimé dans ce modèle et donc doit correspondre à un a priori pur du problème.

Cas Markov : Sun et Cai (2009) ont traité le cas où $H_1, H_2, \dots, H_m$ forment une chaîne de Markov. Dans ce cas, les q-values locales $\ell_i(X)$ peuvent être calculées par l'algorithme forward-backward. L'estimation des hyper-paramètres (π_0, probabilités de transition de la chaîne et Δ) se fait alors par maximum de vraisemblance marginal (approché par l'algorithme EM). Ils montrent que cette approche Bayes empirique fournit une nette amélioration de la procédure BH dans le cas où les paramètres sont effectivement générés par ces lois a priori.

Conclusion

L'approche de Bayes empirique est un outil qui permet de construire facilement des procédures de test multiple qui s'adaptent automatiquement à un certain type de structure d'hypothèses nulles. Notons cependant que le contrôle du risque type I de ces procédures n'est pas toujours garanti, notamment en raison de l'étape "plug-in" qui nécessite des développements théoriques dédiés.

Remerciements

Je tiens à remercier Ismaël Castillo, Catherine Matias et Fanny Villers pour les discussions enrichissantes qui ont contribuées à la rédaction de cette note.

Bibliographie

[1] Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Statist. Soc. Ser. B.
[2] Efron, B. and Tibshirani, R. and Storey, J. and Tusher, V. (2001), Empirical Bayes analysis of a microarray experiment, J. Amer. Statist. Assoc.
[3] Robbins, H. (1956), An empirical Bayes approach to statistics, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.
[4] Storey, J. (2003), The positive false discovery rate : a Bayesian interpretation and the q-value, Ann. Statist.
[5] Sun, W. and Cai, T. (2009), Large-scale multiple testing under dependence, J. Roy. Statist. Soc. Ser. B.