Contrôle du taux de faux positifs dans le cas dépendant bilatéral

Marine Roux [1]

[1] *Domaine Universitaire BP 46, 38402 Saint Martin d'Hères cedex.*
marine.roux@gipsa-lab.grenoble-inp.fr

Résumé. Dans un contexte de test multiple, nous considérons le problème du contrôle du *False Discovery Rate* (FDR) de la procédure de Benjamini-Hochberg (BH) sous une structure de dépendance spécifique. Cette dernière correspond au cas de tests bilatéraux avec des statistiques de tests gaussiennes équi-corrélées. Le résultat principal de notre étude est la démonstration de la conjecture de Reiner-Benaim (2007), qui fournit une borne du FDR dans le cas de deux tests. Par suite, nous étudions également le cas d'un nombre supérieur de tests à l'aide d'une formule exacte.

Mots-clés. Test multiple, FDR, procédure de Benjamini et Hochberg, tests bilatéraux, équi-corrélation, dépendance non positive.

Abstract. In a multiple testing framework, this paper investigates the problem of False Discovery Rate (FDR) control by the Benjamini-Hochberg (BH) procedure for a specific dependence induced by gaussian equi-correlated two-sided tests statistics. The main result of this study is a proof of Reiner-Benaim's conjecture (2007), which provides an upper bound of the FDR for the case of two tests. We also investigate the case of more than two tests by computing FDR via an exact formula.

Keywords. Multiple testing, FDR, Benjamin-Hochberg procedure, two-sided tests, equi-correlation, non-positive dependence.

1 Introduction

1.1 Motivation

L'analyse de données volumineuses incite souvent à se poser un grand nombre de questions simultanément. Il faut alors construire des procédures statistiques capables de répondre pertinemment à ces questions. L'exemple typique se situe en génomique avec les données de puces à ADN, pour lesquelles il s'agit, le plus souvent, d'identifier les gènes différentiellement exprimés entre deux conditions (malades/sains par exemple). Une procédure pertinente serait une procédure qui d'une part sélectionne suffisamment de gènes, et qui d'autre part ne sélectionne pas "trop" de gènes à tort. A cette fin,

une méthode populaire est le contrôle du *False Discovery Rate* (FDR) défini comme la moyenne de la proportion d'erreurs parmi les hypothèses rejetées. En pratique, la méthode la plus utilisée pour contrôler le FDR est la procédure de Benjamini-Hochberg (BH). Si elle garantit le contrôle du FDR sous une dépendance positive (Benjamini et Yekutieli (2001)), son comportement reste toujours assez mal compris dans le cas d'une dépendance non positive. Notre travail apporte des contributions à ce problème en explorant le cas de tests bilatéraux avec des statistiques de tests gaussiennes équi-corrélées.

1.2 Notations et outils

Formellement, nous nous plaçons dans le modèle statistique gaussien équi-corrélé à m et $m_0 \leq m$ fixés,

$$\Big(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), \{\mathcal{N}(\mu\theta, \Sigma(\rho)); (\mu, \rho, \theta) \in \Lambda\}\Big), \tag{1}$$

où $\Lambda = \mathbb{R}_+^* \times [-(m-1)^{-1}, 1] \times \{\theta \in \{0,1\}^m : \#\{k, \theta_k = 0\} = m_0\}$ et $\Sigma(\rho)$ désigne la matrice telle que $\Sigma_{kk}(\rho) = 1$ et $\Sigma_{kl}(\rho) = \rho$ pour $k \neq l$. Nous testons simultanément les hypothèses nulles $H_{0,k}$: "$\theta_k = 0$" contre les hypothèses alternatives $H_{1,k}$: "$\theta_k \neq 0$", à l'aide des statistiques de tests $|X_k|$, $1 \leq k \leq m$. Pour tout $k \in \{1, \ldots, m\}$, les p-valeurs sont données par $p_k = 2(1 - \Phi(|X_k|))$, où Φ est la fonction de répartition de la loi gaussienne centrée réduite.

La procédure de Benjamini et Hochberg (1995) est définie comme suit : soient $p_{(1)} \leq \ldots \leq p_{(m)}$ les p-valeurs ordonnées, résultant des m tests. La procédure BH de niveau q est définie par : $R^{BH} = \{1 \leq k \leq m,\ p_k \leq q\hat{k}/m\}$, où $\hat{k} = \max\{0 \leq k \leq m, p_{(k)} \leq qk/m\}$, avec la convention $p_{(0)} = 0$. Notons que par convention, la procédure BH est identifiée plus haut avec l'ensemble des indices des hypothèses qu'elle rejette *i.e.* pour tout $k \in \{1, \ldots, m\}$, $k \in R^{BH}$ est équivalent à BH rejette $H_{0,k}$.

Nous étudions le FDR de la procédure BH, défini comme la moyenne de la proportion suivante, encore appelée "proportion de fausses découvertes" ou "taux de fausses découvertes" :

$$\mathcal{Q}(R^{BH}) = \frac{|R^{BH} \cap \mathcal{H}_0(\theta)|}{\max(|R^{BH}|, 1)}, \tag{2}$$

où $\mathcal{H}_0(\theta) = \{1 \leq k \leq m, \theta_k = 0\}$ est l'ensemble correspondant aux vraies hypothèses nulles. Nous cherchons ainsi à étudier la quantité suivante, qui ne dépend que de m, m_0, μ et ρ,

$$\mathrm{FDR}(\mu, \rho) = \mathbb{E}[\mathcal{Q}(R^{BH})]. \tag{3}$$

Pour finir, nous introduisons la notion de positive dépendance appelée PRDS (*positively regressively dependent on each one of a subset*), qui est un outil clé du résultat de cette étude.

Définition. (PRDS) Une famille de variables aléatoires réelles $(Y_k)_{1\leq k\leq m}$ est dite PRDS sur un ensemble $S \subset \{1,\dots,m\}$, si pour tout ensemble $D \subset \mathbb{R}^m$ mesurable, croissant [1], la fonction $u \mapsto \mathbb{P}(Y \in D \mid Y_k = u)$ est croissante pour tout $k \in S$.

1.3 Bornes sur le FDR

Au-delà du cadre gaussien, Benjamini et Hochberg (1995) ont montré que la procédure BH contrôle le FDR au niveau qm_0/m *i.e.* FDR $\leq qm_0/m$ lorsque les p-valeurs associées aux vraies hypothèses nulles sont mutuellement indépendantes. De plus, Benjamini et Yekutieli (2001) ont étendu ce contrôle au cas d'une famille de p-valeurs PRDS. Ces derniers proposent également une borne valable quelle que soit la forme de dépendance entre les p-valeurs, qui vaut $\mathcal{B}_q^{BY}(m,m_0) = (qm_0/m)\sum_{k=1}^{m} 1/k$.

En particulier, dans le modèle (1), BH contrôle le FDR au niveau qm_0/m lorsque $\rho = 0$. Pour $\rho \neq 0$, c'est encore le cas pour $m = m_0$ car la famille $(|X_k|)_{1\leq k\leq m}$ est PRDS (Karlin et Rinott (1981), KR81). En revanche, pour $m_0 < m$, la propriété PRDS n'est plus vérifiée. La seule borne validée théoriquement est ainsi celle de Benjamini et Yekutieli. Cependant, les simulations de Reiner-Benaim (2007) ont suggéré que cette dernière pouvait être largement améliorée. Précisément,

Conjecture de Reiner-Benaim. Pour tout $m \geq 2$, pour tout $0 \leq m_0 \leq m$, on a :

$$\sup_{\mu,\rho} \mathrm{FDR}(\mu,\rho) \leq \mathcal{B}_q^{RB}(m,m_0), \qquad \mathcal{B}_q^{RB}(m,m_0) = q\tfrac{m_0}{m}\Big[1+\tfrac{1}{2}\big(1-\tfrac{m_0}{m}\big)\Big]. \tag{4}$$

Nous soulignons que pour $m_0 \in \{1,\dots,m\}$, $\mathcal{B}_q^{BY}/\mathcal{B}_q^{RB} > (2/3)\sum_{k=1}^{m} 1/k$. De plus, comme $\mathcal{B}_q^{RB}(m,m_0) \leq q$, la relation (4) implique en particulier le contrôle du FDR au niveau q.

2 Résultat

Dans cette section, nous présentons une validation théorique de la conjecture de Reiner-Benaim pour $m = 2$.

Théorème. *Pour $m = 2$, la conjecture de Reiner-Benaim est vérifiée.*

Preuve. Le cas $m_0 = 0$ étant trivial, il suffit de prouver le théorème pour $m_0 = 1$ et $m_0 = 2$. Lorsque $m_0 = 2$, la famille des p-valeurs est PRDS (KR81) donc d'après le théorème 1.2 de Benjamini et Yekutieli (2001), $\mathrm{FDR}(\mu,\rho) \leq q$ pour tout $\mu > 0$, pour tout $\rho \in [-1,1]$. Ainsi, il ne reste que le cas $m_0 = 1$. Soient $\mu > 0$ et $\rho \in [-1,1]$, on a,

$$\begin{aligned}\mathrm{FDR}(\mu,\rho) &= \sum_{k=1}^{2}\frac{1}{k}\left[\mathbb{P}\Big(\hat{k}\leq k, p_1 \leq \frac{qk}{m}\Big) - \mathbb{P}\Big(\hat{k}\leq k-1, p_1 \leq \frac{qk}{m}\Big)\right]\\ &= \frac{1}{2}\mathbb{P}\Big(\hat{k}\leq 2, p_1 \leq q\Big) - \frac{1}{2}\mathbb{P}\Big(\hat{k}\leq 1, p_1 \leq q\Big) + \mathbb{P}\Big(\hat{k}\leq 1, p_1 \leq \frac{q}{2}\Big)\end{aligned}$$

1. D $\subset \mathbb{R}^m$ est croissant si pour tout w_1, $w_2 \in \mathbb{R}^m$ tels que pour tout $k \in \{1,\dots,m\}$ $w_{1,k} \leq w_{2,k}$ et $w_1 \in D$ alors $w_2 \in D$.

$$\mathrm{FDR}(\mu,\rho) = \frac{q}{2} - \frac{1}{2}\mathbb{P}\Big(p_2 > q, p_1 \leq q\Big) + \mathbb{P}\Big(p_2 > q, p_1 \leq \frac{q}{2}\Big). \tag{5}$$

On pose $z_1 = \Phi^{-1}(1-\frac{q}{2})$ et $z_2 = \Phi^{-1}(1-\frac{q}{4})$. Notons $X_1 = Y_1$ et $X_2 = Y_2 + \mu$, où Y est un vecteur gaussien de loi $\mathcal{N}(0,\Sigma_2)$. L'expression (5) se réécrit alors,

$$\mathrm{FDR}(\mu,\rho) = \frac{q}{2} - \frac{1}{2}\mathbb{P}(|Y_2+\mu| < z_1, |Y_1| \geq z_1) + \mathbb{P}(|Y_2+\mu| < z_1, |Y_1| \geq z_2).$$

Or $|Y_2+\mu| < z_1 \Leftrightarrow \big[Y_2 \in A \text{ ou } Y_2 \in B\big]$, où $A = \{y \in \mathbb{R}, |y| < (z_1-\mu)_+\}$ et $B = \big]-z_1-\mu, -|z_1-\mu|\big]$. Remarquons que $A \cap B = \varnothing$, A est symétrique par rapport à 0 et $B \subset \mathbb{R}_-$. Par la propriété PRDS du vecteur $|Y|$ sur $\{1,2\}$ (KR81) il vient que $\mathbb{P}(Y_2 \in A, |Y_1| \geq z_2) - \frac{1}{2}\mathbb{P}(Y_2 \in A, |Y_1| \geq z_1) \leq 0$. Ainsi, l'expression (5) se majore par,

$$\begin{aligned}
&\frac{q}{2} - \frac{1}{2}\mathbb{P}(Y_2 \in B, |Y_1| \geq z_1) + \mathbb{P}(Y_2 \in B, |Y_1| \geq z_2) \\
&\leq \frac{q}{2} - \frac{1}{2}\mathbb{P}(Y_2 \in B, |Y_1| \geq z_2) + \mathbb{P}(Y_2 \in B, |Y_1| \geq z_2) \\
&\leq \frac{q}{2} + \frac{1}{2}\mathbb{P}(Y_2 \in B, |Y_1| \geq z_2) \leq \frac{q}{2} + \frac{1}{2}\mathbb{P}(Y_2 < 0, |Y_1| \geq z_2).
\end{aligned}$$

Par symétrie de (Y_1, Y_2), $\mathbb{P}(Y_2 < 0, |Y_1| \geq z_2) = \frac{1}{2}\mathbb{P}(|Y_1| \geq z_2) = \frac{q}{4}$. Finalement, $\mathrm{FDR}(\mu,\rho) \leq \frac{5q}{8} = \mathcal{B}_q^{RB}(2,1)$. □

La figure 1 illustre le théorème précédent en utilisant une formule exacte du FDR, établie de la même façon que dans la proposition 3.8 de Roquain et Villers (2011).

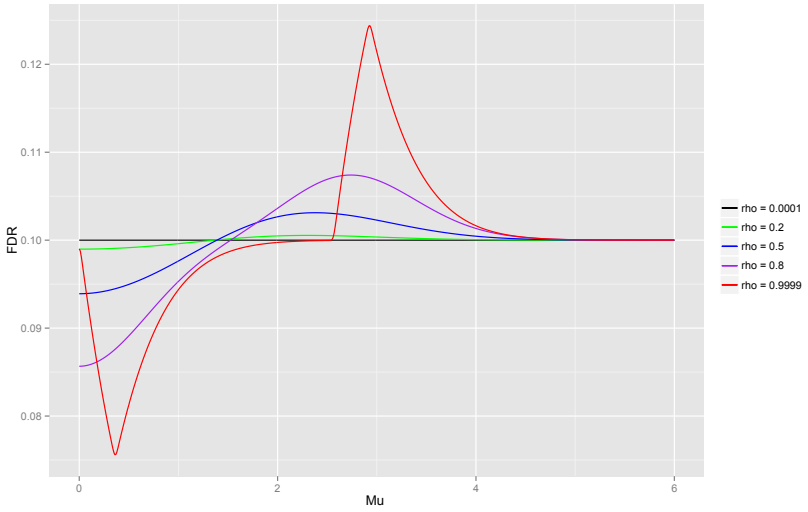
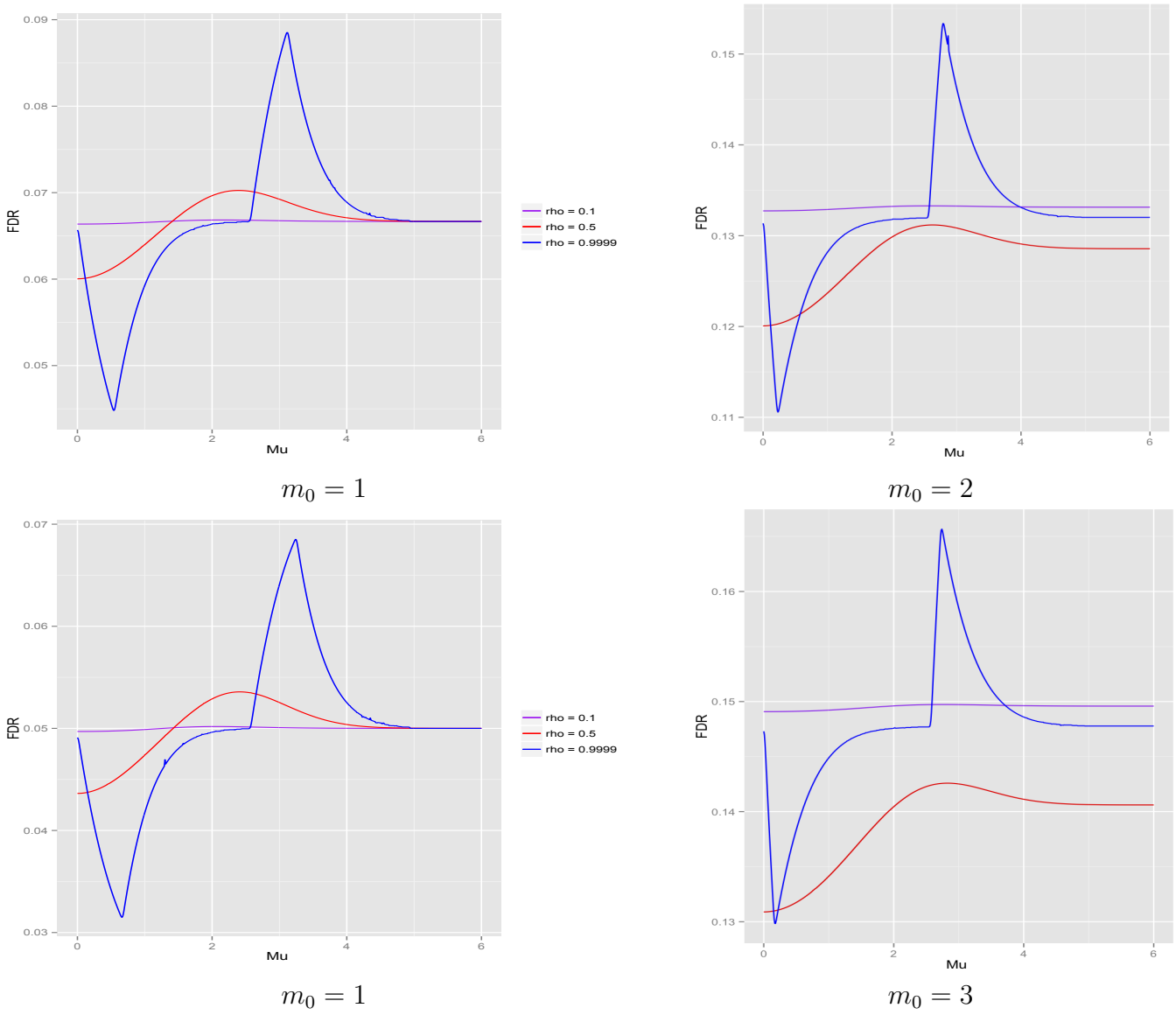


FIGURE 1 – FDR (donné par (3)) en fonction de μ pour différentes valeurs de ρ, $m_0 = 1$, $m = 2$, $q = 0.2$.

3 Extension

Dans cette section, nous explorons la conjecture de Reiner-Benaim pour certaines valeurs de $m \geq 3$. En adaptant au contexte bilatéral le théorème 3 de Roquain et Villers (2011), nous avons obtenu une formule exacte du FDR valable pour $m \geq 3$ et $\rho > 0$. Cette expression a permis de fournir la figure 2 :



$m_0 = 1$ $\quad$ $m_0 = 2$

$m_0 = 1$ $\quad$ $m_0 = 3$

FIGURE 2 – FDR(μ, ρ) en fonction de μ pour différentes valeurs de ρ, $m = 3$ (en haut) et $m = 4$ (en bas), $q = 0.2$.

La figure 2 suggère que la borne (4) est valable pour $m = 3$ et $m = 4$. Pour un nombre de tests plus important, les conclusions sont similaires dans les cas traités ($m \leq 20$).

Conclusion. Notre étude a contribué au développement de la théorie du contrôle du FDR dans un cas de dépendance non positive, issue du cas gaussien équi-corrélé. La suite naturelle de cette étude serait la démonstration de la conjecture de Reiner-Benaim pour $m \geq 3$. Par ailleurs, les simulations semblent indiquer que la borne $\mathcal{B}_q^{RB}(m, m_0)$ est encore valable pour un type de corrélation à décroissance exponentielle. Ainsi, il serait intéressant d'identifier précisément les différents types de matrices de covariance pour lesquels BH contrôle le FDR au niveau q.

Remerciements

Cette étude a été en partie réalisée lors de mon stage de master. Je remercie Sophie Achard (GIPSA-Lab, Grenoble), Irène Gannaz (INSA de Lyon) et Etienne Roquain (UMPC, Paris 6) pour leurs encadrements et leurs collaborations. Je remercie Vincent Rivoirard (Université Paris-Dauphine) qui m'a permis d'étoffer cette étude après mon stage. Enfin, je remercie Sophie Achard, Pierre Borgnat (ENS Lyon), Irène Gannaz et Etienne Roquain pour leurs accompagnements actuels en thèse. Ce travail a été financé par l'institut Camille Jordan (Lyon 1), le GIPSA-Lab (Grenoble), l'Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration et ANR-14-CE27-0001).

Bibliographie

[1] Benjamini, Y. et Hochberg Y. (1995), *Controlling the False Discovery Rate : A Pratical and Powerful Approach to Multiple Testing*, Royal Statistical Society, 57, 289–300.
[2] Reiner-Benaim, A. (2007), *Control by the BH Procedure for Two-Sided Correlated Tests with Implications to Gene Expression Data Analysis*, Biometrical Journal, 49, 107-126.
[3] Benjamini, Y. et Yekutieli, D. (2001), *The control of the false discovery rate in multiple testing under dependency*, The Annals of Statistics, 29, 1165-1188.
[4] Karlin, S. et Rinott, Y. (1981), *Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities* , The Annals of Statistics, 9, 1035-1049.
[5] Roquain, E. et Villers, F. (2011), *Exact calculation for false discovery proportion with application to least favorable configurations*, The Annals of Statistics, 39, 1 584-612.