Contrôle des erreurs pour la détection d'événements rares et faibles dans des champs de données massifs

Céline Meillier, Florent Chatelain, Olivier Michel, Hacheme Ayasso

GIPSA-Lab, 11 rue des Mathématiques, Grenoble Campus, BP46, F-38402 Saint Martin d'Hères Cedex
prenom.nom@gipsa-lab.fr

Résumé − Nous nous intéressons à la détection d'événements rares et de faible intensité dans des données massives bruitées. Les approches par tests multiples d'hypothèses peuvent être utilisées pour extraire une liste d'échantillons susceptibles de contenir de l'information tout en contrôlant un critère d'erreurs de détection global. Dans la littérature, la plupart de ces approches ne sont valides que pour des tests indépendants, ou sous des hypothèses particulières de dépendance. Nous nous proposons de montrer, en étendant les travaux de Benjamini et Yekutieli [1], que sous certaines hypothèses, il est cependant possible d'appliquer la procédure de contrôle du taux de fausses découverte (FDR) de Benjamini-Hochberg [2] sur une statistique de filtrage adapté très utilisée en traitement du signal et des images.

Abstract − In this paper, we address the general issue of detecting rare and weak signatures in very noisy data. Multiple hypotheses testing approaches can be used to extract a list of components of the data that are likely to be contaminated by a source while controlling a global error criterion. However most of efficients methods available in the literature stand for independent tests, or require specific dependency hypotheses. Based on the work of Benjamini and Yekutieli [1], we show that under some classical positivity assumptions, the Benjamini-Hochberg procedure for False Discovery Rate (FDR) [2] control can be directly applied to the statistics produced by a very common tool in signal and image processing that introduces dependency: the matched filter.

1 Introduction

Considérons l'observation d'événements rares et de faible intensité. Nous supposons que la contribution de ces événements dans le signal se restreint à une faible proportion des échantillons du signal, le reste se résumant à un bruit blanc gaussien. C'est par exemple le cas d'objets non résolus dans des grandes images ou encore de signaux à durée très limitée dans une série temporelle. Ces événements seront appelés sources dans la suite de ce papier. Nous souhaitons détecter le plus grand nombre d'échantillons susceptibles d'appartenir à une source tout en limitant le nombre de *fausses découvertes*, *i.e* le nombre d'échantillons qui ne contiennent que du bruit mais qui ont été détectés comme contribution significative d'une source. Nous considérons le modèle d'observation linéaire suivant :

$$Y = H\boldsymbol{a} + \epsilon, \tag{1}$$

où $Y \in \mathbb{R}^d$ est le vecteur d'observation (pour une image, Y est la version vectorisée de l'image), $\epsilon \in \mathbb{R}^d$ est un bruit blanc gaussien (bruit de mesure, de l'environnement, etc). $H \in \mathbb{R}^{d\times n}$ est une matrice de régression, un dictionnaire de sources ou de signaux élémentaires par exemple, où n est la dimension de l'espace des solutions possibles, et $\boldsymbol{a} \in \mathbb{R}^n$ est le vecteur des intensités. Dans le cas d'événement rares, le nombre de coefficients a_i non nuls de $\boldsymbol{a}$ doit être faible devant la dimension d des observations et devant n.

Les méthodes de sélection de modèles, largement utilisées pour traiter ce problème de détection, reposent souvent sur la minimisation de critères des moindres carrés pénalisés $\widehat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} ||Y - H\boldsymbol{a}||_2^2 + \lambda\phi(\boldsymbol{a})$ où la pénalisation $\phi(\boldsymbol{a})$ est choisie de manière à favoriser la parcimonie de la solution (par exemple, la norme ℓ_1, ou la pseudo norme ℓ_0 du type critère AIC, etc). Le vecteur des intensités $\boldsymbol{a}$ est ainsi seuillé sans contrôle global des erreurs. Nous souhaitons ici obtenir un tel contrôle fournissant un critère interprétable en terme de fausses détections.

Décider quels échantillons appartiennent à une source ou ne contiennent que du bruit peut aussi se formuler sous la forme de tests multiples pour tout $1 \leq i \leq n$:

$$\left\{ \begin{array}{lcll} \mathcal{H}_0^i & : & a_i = 0 & \text{(bruit seul)} \\ \mathcal{H}_1^i & : & a_i > 0 & \text{(source + bruit)} \end{array} \right. \tag{2}$$

Seuiller le vecteur d'intensité équivaut à rejeter les hypothèses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^n$ selon un certain critère d'erreur. Si tous les tests sont seuillés avec un même niveau de significativité α (défini pour un seul test), le nombre n (très grand) de tests n'est pas pris en compte. Le nombre moyen de fausses alarmes est alors contrôlé au niveau $n\alpha$ ce qui peut conduire à une très grande proportion de fausses alarmes parmi tous les tests rejetés. Afin de contrôler la probabilité α de faire au moins une fausse alarme, une correction de type Bonferroni [3] peut être appliquée en seuillant chaque

test individuel à un niveau α/n. Cette procédure est très conservative et la plupart des échantillons appartenant à une source ne seront pas détectés. Benjamini et Hochberg [2] ont proposé une méthode plus puissante permettant de contrôler un taux d'erreur global : le taux de fausses découvertes (FDR), *i.e* le taux de vraies $\mathcal{H}_0$ rejetées à tort parmi toutes les hypothèses rejetées. Le contrôle du FDR avec la procédure de Benjamini-Hochberg (BH) a été largement appliqué à différents domaines : l'astronomie [4], la neuroimagerie [5] ou la génomique [6].

Dans le cas d'événements de faible intensité noyés dans du bruit, il est nécessaire de prétraiter les données pour améliorer la détection. Un outil classique lorsque la signature des événements est connue est le filtrage adapté afin de maximiser le rapport signal à bruit (RSB) entre les contributions des sources et celle du bruit. Or ce filtrage introduit des corrélations entre les échantillons. Quelques rares procédures permettant de contrôler le FDR dans certains cas de dépendance ont été proposées dans la littérature. Récemment dans [7] les auteurs proposent une méthode de sélection de variables, appelée le *knockoff filter*, qui contrôle le FDR dans le cas du modèle linéaire eq. (1). Cependant dans les cas qui nous intéressent, les fortes corrélations locales et le nombre massif de tests interdisent de construire ces *knockoffs*. Dans [1], un facteur correctif est ajouté dans la procédure BH afin de conserver le contrôle quelle que soit la structure de dépendance ; cette méthode s'avère bien trop conservative pour être utilisée. Les auteurs montrent également qu'il est possible de contrôler le FDR dans le cas de dépendance avec la procédure originale BH sous certaines conditions de positivité. Nous nous appuierons sur les résultats de [1] pour montrer qu'il est possible de contrôler le FDR en seuillant le résultat d'un filtrage adapté avec la procédure BH sous réserve de satisfaire quelques conditions réalistes dans le cadre de la détection de sources.

2 Formulation du problème

2.1 Modèle d'observation et hypothèses

Nous décrirons dans la suite du papier les observations et la signature des sources sous forme vectorisée, y compris dans le cas de signaux en deux dimensions ou plus. Les observations $Y \in \mathbb{R}^d$ peuvent se décomposer ainsi :

$$Y = \sum_j s_j + \epsilon \tag{3}$$

où $s_j \in \mathbb{R}^d$ est la réponse de l'instrument de mesure à la $j^{ème}$ source observée et $\epsilon \in \mathbb{R}^d$ est un bruit gaussien : $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$. La variance est supposée connue, et sans perte de généralité on pose $\sigma^2 = 1$. La réponse de la $j^{ème}$ source peut s'écrire : $s_j(r) = \sum_i a_i^j h(r, r_i)$, où i décrit le support de la source, a_i^j est la contribution de la source à la position r_i et $h(\cdot, r_i)$ est la réponse du système (normalisée ℓ_2) au point r_i. Cette réponse étant supposée à support limité, les événements observés étant rares et également à support limité, la plupart des composantes du vecteur observé Y ne contiennent que du bruit. Le vecteur d'intensité $\boldsymbol{a} = [a_1, \cdots, a_n]^T$ de l'éq. (1) est défini par ses coefficients $a_i = \sum_j a_i^j$ obtenus en sommant les contributions des différentes sources s_j à la position r_i.

Nous supposons maintenant les deux hypothèses suivantes : 1) la contribution de la source est positive, $a_i^j > 0$ pour tout $1 \leqslant i \leqslant m$ où m est le nombre d'échantillons du support de la source et 2) la réponse du système est non négative, pour tout r et tout r_i : $h(r, r_i) \geqslant 0$.

2.2 Détection

Le nombre de sources et leur position sont *a priori* inconnus ; il faut donc tester les d positions possibles dans le champ de données. Le vecteur observé Y s'exprime selon l'éq. (1) dans laquelle H sera une matrice de taille $d \times n$ dont chaque colonne représente la réponse $h(\cdot, r_i)$ centrée sur la position r_i, $1 \leqslant i \leqslant n$. Dans ce cas, $n = d$ puisqu'il faut tester chaque position possible dans les données. Tester la présence de sources à la position r_i revient à tester la valeur a_i avec les hypothèses définies par l'eq. (2).

Afin d'améliorer la détectabilité des sources, nous effectuons un filtrage adapté à la réponse du système qui s'écrit :

$$H^T Y = H^T H \boldsymbol{a} + H^T \epsilon \tag{4}$$

La matrice H présente des propriétés importantes pour la suite du processus de détection :

1. H est creuse, *i.e* elle contient peu de coefficients significatifs.
2. $(H^T H)_{i,i} = 1 \; \forall \; 1 \leqslant i \leqslant n$.
3. $H \geqslant 0$ est une matrice non négative, *i.e.* toutes les composantes de la matrice sont non négatives.

D'après l'équation (4), le vecteur $H^T Y$ est gaussien :

$$H^T Y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5}$$

où $\boldsymbol{\mu} = H^T H \boldsymbol{a} \geqslant 0$ et $\boldsymbol{\Sigma} = H^T H \geqslant 0$.

3 Seuillage et contrôle du FDR

Le filtrage adapté modifie la formulation des tests décrits par l'éq. (2), un plus grand nombre d'échantillons vont être classés dans $\mathcal{H}_1$ car le filtrage adapté élargit les sources. Notons $X = H^T Y$ le vecteur de $\mathbb{R}^d$ résultant du filtrage adapté. D'après l'éq. (5), chaque composante X_i de X est gaussienne : $X_i \sim \mathcal{N}(\mu_i, 1)$, pour tout $1 \leqslant i \leqslant n$. La détection des contributions des sources dans la nouvelle statistique X conduit alors au test suivant

$$\begin{cases} \mathcal{H}_0^i & : \quad \mu_i = 0 \quad \text{(bruit seul)} \\ \mathcal{H}_1^i & : \quad \mu_i > 0 \quad \text{(contribution d'une source)}, \end{cases} \tag{6}$$

Dans le cas de champs de données massifs, le nombre de tests n peut être très important, il est indispensable de contrôler un critère d'erreur global.

3.1 P-valeurs et procédure BH

Dans le cas de n tests indépendants, la procédure BH proposée dans [2] contrôle le FDR à un niveau $\pi_0 q$ où $\pi_0 = \frac{n_0}{n}$ et n_0 est le nombre de tests réellement sous $\mathcal{H}_0$ et $0 \leq q \leq 1$ est le paramètre de contrôle. Rappelons brièvement la procédure BH :

1. Soient $p_{(0)} < p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(n)}$ les n p-valeurs ordonnées, avec par convention $p_{(0)} = 0$.
2. Soit $k = \underset{i}{\text{argmax}} \left(p_{(i)} \leqslant q\frac{i}{n} \right)$
3. Rejet des hypothèses $\mathcal{H}_0^{(1)}, \cdots, \mathcal{H}_0^{(k)}$.

Dans le problème défini par l'éq. (6), les p-valeurs sont calculées à partir de la fonction de répartition $F_{\mathcal{H}_0}$ de la loi de X sous $\mathcal{H}_0$ (loi normale standard) : $p_i = 1 - F_{\mathcal{H}_0}(X_i)$. Par construction les p-valeurs p_i sont uniformément distribuées sur $[0, 1]$ sous $\mathcal{H}_0$. La fonction de répartition $F_{\mathcal{H}_0}$ est stochastiquement plus grande que celle sous $\mathcal{H}_1$, $F_{\mathcal{H}_0}(x) \geqslant F_{\mathcal{H}_1}(x)$.

3.2 Filtrage adapté et contrôle du FDR

D'après les travaux de Benjamini et Yekutieli [1], si la distribution des p-valeurs est PRDS (pour *positive regression dependency on a subset*[1]) alors la procédure BH contrôle le FDR à un niveau inférieur ou égal à $\pi_0 q$. De plus, lorsque $\mathbf{\Sigma} \geqslant 0$ alors $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ est PRDS, et si f est une fonction monotone alors le vecteur $Y = (f(X_1), \cdots, f(X_n))$ est aussi PRDS. Nous pouvons donc écrire :

Proposition. *La procédure de Benjamini-Hochberg permet de seuiller la sortie du filtrage adapté tout en contrôlant le taux de fausses découvertes (FDR).*

Dans le cas qui nous intéresse, le résultat du filtrage adapté $X = H^T Y$ est en effet PRDS puisqu'il est gaussien avec une matrice de covariance non-négative, voir éq. (5). Les p-valeurs se déduisant par une fonction décroissante du vecteur gaussien X sont donc également PRDS.

La procédure BH permet de contrôler le FDR au niveau $q\pi_0 \leqslant q$ en seuillant les p-valeurs correspondantes. Notons que si π_0 n'est pas connue, le contrôle se fait au niveau q. Finalement, rejeter $\mathcal{H}_0^{(1)}, \cdots, \mathcal{H}_0^{(k)}$ est équivalent à dire que les composantes correspondantes, $X_{(1)}, X_{(2)}, \cdots, X_{(k)}$, sont significatives, *i.e* $\mu_{(1)} > 0, \mu_{(2)} > 0, \cdots, \mu_{(k)} > 0$.

3.3 Validation sur données simulées

Les performances du seuillage par la procédure BH sont mesurées sur une image synthétique de 150×150 pixels (soient $n = d = 22500$ tests à considérer) contenant douze sources non résolues avec le même rapport signal à bruit, voir figures 1(a) et 1(b). La PSF simulée est une fonction gaussienne 2D de largeur à mi-hauteur 3,5 pixels. Dans ce cas simulé, la proportion π_0 est connue, elle vaut $\pi_0 = 0,85$. Nous constatons que des seuillages de chaque test à différents niveaux α (figure 1(e)) entraînent de très nombreuses fausses détections. A l'inverse la procédure BH (figure 1(f)) limite grandement les fausses détections tout en gardant un très bon taux de détection. Ces perfor-

1. Le lecteur peut se reporter à [1, p1168] pour une définition complète.

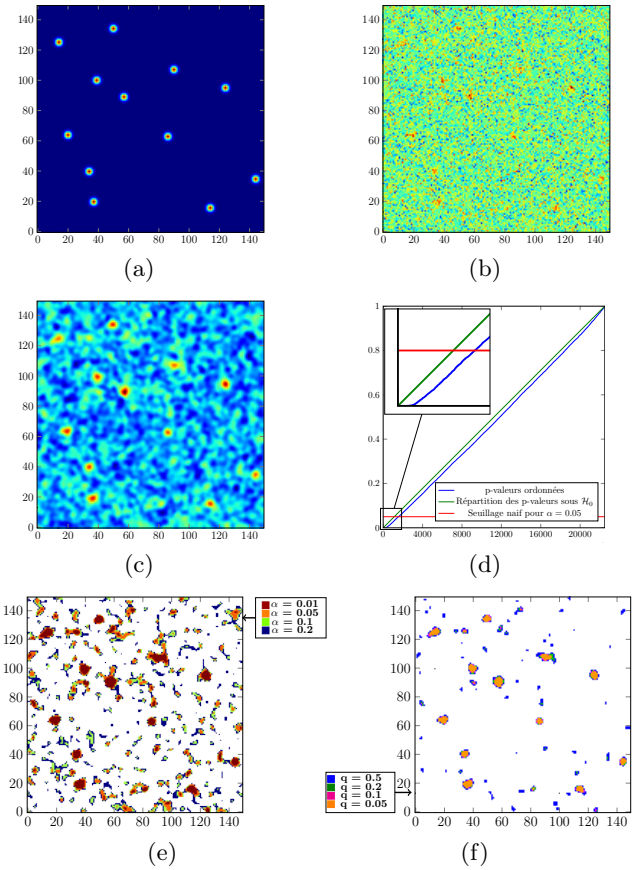

FIGURE 1 – Exemple de recherche de sources rares et faibles : (a) Réponses des douze sources, (b) observation bruitée (RSB = -5dB), (c) sortie du filtrage adapté, (d) p-valeurs tracées en fonction de leur rang (courbe bleue), quantiles théoriques i/n sous l'hypothèse nulle en fonction de i (courbe verte : $y = i/n$), seuillage individuel des tests pour $\alpha = 0.05$ (courbe rouge), (e) superposition du seuillage individuel des tests pour différentes valeurs de α, et (f) superposition du seuillage par la procédure BH pour différents niveaux de contrôle q du FDR.

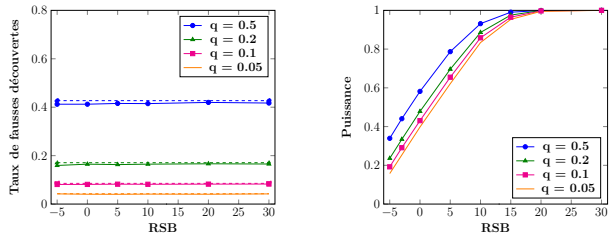

FIGURE 2 – Taux de fausses découvertes et niveau de contrôle théorique $\pi_0 q$, resp. en trait plein et pointillés (a), et puissance en fonction du RSB pour différents niveaux q de contrôle du FDR (b).

mances estimées sur 1000 simulations de l'image bruitée de douze sources sont montrées sur la figure 2 en fonction du RSB. Le FDR obtenu est toujours proche du niveau de contrôle alors que la puissance augmente avec le RSB.

4 Détection de galaxies

Dans le cadre de l'application astrophysique appliquée aux données MUSE (Multi Unit Spectroscopic Explorer), la procédure présentée dans la section 3 permet de sélectionner les ensembles de pixels susceptibles de contenir des sources intéressantes. Elle permet d'établir une carte de proposition qui peut être exploitée par des algorithmes de détection de sources comme la méthode présentée dans [8] afin de réduire l'espace d'exploration des données.

4.1 Données MUSE et objectifs

MUSE est un instrument destiné à l'observation des galaxies lointaines. Il produit des données hyperspectrales composées de 3600 images de 300×300 pixels pour des longueurs d'onde $\lambda \in [480\text{nm}, 930\text{nm}]$. Nous travaillerons ici avec les observations MUSE du champ de données HDFS [2] (Hubble Deep Field South). Les galaxies lointaines que nous cherchons à détecter sont faiblement résolues spatialement et de faible intensité. Leur spectre se réduit à une raie d'émission large de quelques longueurs d'onde. La position des galaxies et la position de leur raie sont inconnues. La réponse h de l'instrument (PSF) étant mesurée et modélisée, nous appliquons un filtrage adapté à cette PSF afin d'améliorer le RSB des galaxies lointaines. Dans le cas de MUSE, h varie spectralement. La matrice H est alors obtenue à partir de l'expression de h pour les quelques $n = d \approx 3 \times 10^8$ positions possibles.

4.2 Seuillage des données 3D par la procédure BH

Dans le cas de MUSE, une carte de variance empirique S^2, indépendante de Y et dont les composantes sont indépendantes et distribuées selon une loi du χ^2, est fournie avec les données. Le vecteur normalisé X/S suit une loi de Student ; il n'est plus PRDS mais la procédure BH contrôle toujours le FDR à condition que $q < 0.5$ [1], ce qui est le cas en pratique. Le seuillage du cube met en évidence sur la figure 3 un grand nombre de régions de faible extension dans les domaines spatial et spectral, rejetées par la procédure BH (environ 9% du cube pour un contrôle à $q = 0.1$). Tous ces pixels forment la carte de proposition pondérée exploitable par l'algorithme de détection de galaxies présenté dans [8], ce qui permet de réduire le temps de calcul de façon importante. Cette approche présente l'avantage de fournir un critère de contrôle global des erreurs dans la carte de proposition contrairement au max-test mis en oeuvre dans [10] et [8] où nous contrôlons un taux de fausses alarmes pour chaque spectre individuellement.

2. Données disponible sur le site http ://muse-vlt.eu/science. Voir [9] pour une description détaillée des données HDFS.

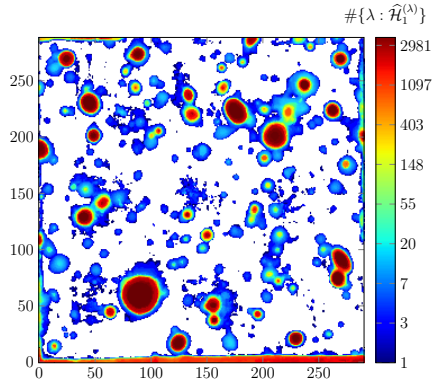

FIGURE 3 – Somme selon l'axe des longueurs d'onde du cube après seuillage binaire. L'échelle de couleur indique pour chaque spectre (*i.e.* chaque position (x, y)) le nombre de découvertes (les $\mathcal{H}_0$ rejetées) par la procédure BH pour un contrôle du FDR $q = 0.1$.

5 Conclusion

Sous des hypothèses de positivité, nous avons montré qu'il est possible de seuiller la statistique du filtre adapté avec la procédure de Benjamini-Hochberg tout en contrôlant le FDR. Notons que si la réponse du système h contient quelques valeurs négatives, il est envisageable de réaliser un filtrage adapté sous-optimal avec la version tronquée h^+ de h telle $h^+(r) = h(r)$ si $h(r) > 0$ sinon $h^+(r) = 0$. Le contrôle du FDR pour construire une carte de proposition est particulièrement intéressant dans le cadre de l'exploration de données massives afin de réduire les temps de calcul inhérent à leurs dimensions.

Références

[1] Y. Benjamini, D. Yekutieli. *The control of the false discovery rate in multiple testing under dependency.* Annals of statistics, 2001, p1165-1188.

[2] Y. Benjamini, Y. Hochberg. *Controlling the false discovery rate : a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, Series B (Methodological), 1995, p289-300.

[3] S. Holm. *A simple sequentially rejective multiple test procedure.* Scandinavian journal of statistics, 1979, p65-70.

[4] C.J. Miller, et al. *Controlling the false discovery rate in astrophysical data analysis.* The Astronomical Journal, 2001, p3492.

[5] C. R. Genovese, et al. *Thresholding of statistical maps in functional neuroimaging using the false discovery rate.* Neuroimage, 2002, p870-878.

[6] A. Reiner, et al. *Identifying differentially expressed genes using false discovery rate controlling procedures.* Bioinformatics, 2003, p368-375.

[7] R. F. Barber, E. Candes. *Controlling the false discovery rate via knockoffs* arXiv preprint arXiv :1404.5609, 2014.

[8] C. Meillier, et al. *Nonparametric bayesian extraction of object configurations in massive data.* IEEE TSP, 2015 .

[9] R. Bacon, et al. *The MUSE 3D view of the Hubble Deep Field South.* Astronomy and Astrophysics, 2015

[10] S. Paris, et al. *Detection tests using sparse models, with application to hyperspectral data.* IEEE TSP, 2013 .