Méthode de σ-clipping par point fixe pour l'estimation de la distribution sous $\mathcal{H}_0$ dans le cadre de tests multiples

Céline MEILLIER[1], Raphael BACHER [2], Florent CHATELAIN[2], Olivier MICHEL[2]

[1]Laboratoire Icube
300 boulevard Sebastien Brant, 67400 Illkirch-Graffenstaden

[2]GIPSA-lab
11 rue des mathématiques, 38400 Saint Martin d'Hères

meillier@unistra.fr, raphael.bacher@gipsa-lab.fr
florent.chatelain@gipsa-lab.fr, olivier.michel@gipsa-lab.fr

Résumé – Dans le cadre des procédures de type tests multiples (imagerie médicale, astronomie, données génétiques, etc), il est nécessaire de modéliser l'hypothèse nulle $\mathcal{H}_0$ qui représente le cas où "le phénomène d'intérêt est absent", *i.e.* il n'y a que du bruit. De nombreux travaux [3, 4, 5] ont mis en évidence le fait que, même sous $\mathcal{H}_0$, les données traitées ne sont pas exactement distribuées selon le modèle théorique utilisé. Il s'avère alors nécessaire d'estimer la distribution sous $\mathcal{H}_0$ à partir des données elles-même afin de corriger le modèle et d'obtenir des procédures de test plus fiables et plus puissantes. Une telle estimation doit être robuste face à la présence potentielle du phénomène/signal d'intérêt (hypothèse alternative $\mathcal{H}_1$) parmi les multiples observations. Nous proposons dans ce papier une nouvelle méthode d'estimation des paramètres de moyenne et de variance sous l'hypothèse nulle $\mathcal{H}_0$. Nous montrons une comparaison des performances par rapport à des méthodes issues de la littérature dans le cas de champs aléatoires gaussiens dont la moyenne et la variance sont *a priori* inconnues.

Abstract – Multiple tests procedures used in different fields such as medical imaging, astronomy, or genetics require the knowledge of the null distribution under $\mathcal{H}_0$ which corresponds to the case where the phenomenon of interest is absent, *i.e.* there is only noise. Some authors as [3, 4, 5] highlight the fact that, even the true $\mathcal{H}_0$ samples are not exactly distributed as expected. It is then necessary to estimate the true null distribution directly from the data to correct this misspecification. This ensures more reliable and powerful testing procedures, robust to the possible presence of the signal/phenomenon of interest ($\mathcal{H}_1$ alternative hypothesis). In this paper, we propose a new method for mean and variance estimation under $\mathcal{H}_0$. We show a comparison of the performances w.r.t. well-known methods of the literature in the case of random gaussian field data whose mean and variance are *a priori* unknown.

1 Introduction

Considérons l'observation de signaux de faible intensité noyés dans du bruit. Nous supposons que la proportion d'échantillons ne contenant pas de signal est majoritaire. Plus le rapport signal à bruit (RSB) des signaux est faible plus il est difficile de séparer les échantillons de bruit seul et les échantillons contenant du signal utile. La prise de décision peut se formuler sous la forme de tests d'hypothèses binaires :

$$\begin{cases} \mathcal{H}_0 & : \text{ bruit seul} \\ \mathcal{H}_1 & : \text{ signal + bruit} \end{cases}$$

Dans le cas de tests multiples, chaque échantillon est testé pour les hypothèses $\mathcal{H}_0$/$\mathcal{H}_1$ et l'on peut appliquer une procédure permettant de contrôler un critère d'erreur global, par exemple le taux de fausses découvertes (FDR) [2] ou le "Higher Criticism" [1]. Ces procédures nécessitent alors de connaître la loi des échantillons, ou plus généralement des statistiques de test, sous $\mathcal{H}_0$. Dans de nombreux cas, on utilise un modèle paramétrique pour cette distribution, par exemple une loi gaussienne de moyenne et variance théoriques connues. Dans la pratique, les données sous $\mathcal{H}_0$, sont rarement parfaitement distribuées selon le modèle utilisé. Ceci pour diverses raisons telles que l'effet de prétraitements, un modèle trop éloigné de la réalité, des statistiques de test calculées sous des hypothèses d'indépendance peu réalistes, etc. Le problème dû à la différence entre la distribution sous $\mathcal{H}_0$ théorique (modèle) et la distribution réelle des données a été relevé dans différents travaux [3, 4, 5]. Dans un contexte de tests multiples où un grand nombre d'échantillons sont disponibles, il devient néanmoins possible de corriger cette mauvaise spécification de l'hypothèse nulle. La principale difficulté réside dans le fait qu'une partie, inconnue, des données observées est distribuée sous l'hypothèse alternative $\mathcal{H}_1$ (présence d'un signal utile dans notre problème de détection). Il s'agit alors d'un problème d'estimation robuste afin de ne pas être significativement biaisé par ces échantillons "aberrants". En estimation robuste, on note trois catégories de méthodes répondant à notre problème. On peut i) modifier le modèle statistique du bruit en utilisant une distribution à queue lourde qui permet de modéliser l'incertitude dans la queue de la distribution due aux échantillons sous $\mathcal{H}_1$, ii) modifier la fonc-

tion objectif à optimiser afin que la solution soit peu biaisée par les valeurs extrêmes (M-estimateurs, pénalisation de Huber, voir [6]), iii) réaliser l'estimation à partir de statistiques d'ordre (L-estimateur, voir [6]) ou sur des données tronquées. C'est dans cette dernière catégorie que se situent les méthodes [3, 4, 5] et le présent travail. Dans le cadre de tests à grande échelle, une troncature des valeurs extrêmes permet en effet une estimation très robuste (biais très faible) tout en conservant suffisamment de données pour garantir une faible variance de l'estimateur.

Dans la suite, nous supposerons que les échantillons sont distribués selon une loi gaussienne $\mathcal{N}(\mu, \sigma)$ sous $\mathcal{H}_0$, i.e. que le bruit est gaussien pour notre problème de détection. Il s'agit alors d'estimer la médiane μ (équivalente à la moyenne par symétrie de la loi gaussienne) et l'écart-type σ de la loi des échantillons sous $\mathcal{H}_0$. Notons que cette hypothèse est classique pour des problèmes de détection. Plus généralement, de nombreux tests statistiques reposent directement sur un z-score qui suit, au moins approximativement, une loi normale sous $\mathcal{H}_0$ (tests de de Student ou Welch pour de grands échantillons,...). Enfin la méthode proposée peut être aisément élargie à des familles de lois symétriques telles que la loi de Student.

Le papier est organisé de la façon suivante : dans la partie 2 nous présentons la méthode d'estimation de la moyenne et de la variance reposant sur un principe de σ-clipping selon un problème de point fixe, tandis que dans la partie 3 nous comparons les performances de cette nouvelle méthode à celles de méthodes de la littérature. Les résultats montrés ici font partie d'une étude préliminaire qui sera étendue et développée dans une version longue de ces travaux.

2 Procédure de σ-clipping par point fixe

Soit un ensemble de N échantillons $I = \{x_1, ..., x_N\}$ dont une proportion π_0 sont générés selon la loi des données sous $\mathcal{H}_0$, ici la loi $\mathcal{N}(\mu, \sigma)$, dont on note F_0 la fonction de répartition, et une proportion $\pi_1 = 1 - \pi_0$ sont générés selon la fonction de répartition F_1 associée à l'hypothèse alternative $\mathcal{H}_1$. La distribution globale des données peut alors s'écrire comme un modèle de mélange à deux groupes de fonction de répartition

$$F(x) = \pi_0 F_0(x) + \pi_1 F_1(x). \tag{1}$$

Le principe du σ-clipping est de tronquer les données sur un domaine centré autour de la médiane et de largeur proportionnelle à l'écart-type σ. L'objectif est d'écarter les échantillons "aberrants", i.e. ceux distribués selon la loi F_1, afin de ne pas biaiser l'estimation de μ et σ. La fonction de répartition des données tronquées F_t, où l'indice t fait référence à la troncature, s'écrit :

$$F_t(x) = \frac{F(x) - F(l)}{F(r) - F(l)} \tag{2}$$

où l (resp. r) correspond au seuil de la troncature à gauche (resp. à droite), défini respectivement comme :

$$l = \mu - \kappa\sigma, \quad r = \mu + \kappa\sigma. \tag{3}$$

Les quartiles $q_{i,t}$, définis comme $F_t\left(q_{i,t}\right) = \frac{i}{4},$ pour $i = 1, 2, 3$, permettent en outre d'obtenir des mesures robustes de location et de dispersion de cette distribution. D'après (2), ces quartiles s'expriment à partir de la distribution non tronquée comme

$$q_{i,t} = F^{-1}\left(\frac{i}{4}F(r) + \left(1 - \frac{i}{4}\right)F(l)\right). \tag{4}$$

La méthode d'estimation repose sur le postulat suivant :

P1. La probabilité d'observer un échantillon distribué selon $\mathcal{H}_1$ est nulle sur le domaine de troncature : $F_1(r) = F_1(l)$.

Cette hypothèse est bien sûr une version idéalisée de la réalité. Néanmoins, elle s'avère une bonne approximation dans les situations classiques de tests multiples comme expliqué dans [8] dans un contexte voisin, en attestent ici les simulations conduites au paragraphe 3.

A partir de P1, il devient possible d'exprimer les quartiles de la loi tronquée $q_{i,t}$ en fonction de la distribution F_0 des données sous $\mathcal{H}_0$:

$$\frac{i}{4} = F_t(q_{i,t}) = \frac{F_0(q_{i,t}) - F_0(l)}{F_0(r) - F_0(l)} = \frac{F_0(q_{i,t}) - F_0(l)}{1 - 2F_0(l)}, \tag{5}$$

pour $i = 1, 2, 3$, où la première égalité vient de la définition des quartiles, la seconde vient de (1) et du postulat P1, et la dernière vient de $F_0(r) = 1 - F_0(l)$ par symétrie des bornes définies en (3) et de F_0 (fonction de répartition de la loi $\mathcal{N}(\mu, \sigma)$).

2.1 Point fixe pour la médiane μ

La médiane des données tronquées correspond au second quartile $q_{2,t}$. D'après (5), il vient que $F_0(q_{2,t}) = \frac{1}{2}$ ce qui montre que $q_{2,t}$ est également la médiane de F_0, i.e. $q_{2,t} = \mu$. Par conséquent, on obtient d'après (4) l'équation au point fixe

$$\mu = F^{-1}\left(\frac{F(r) + F(l)}{2}\right), \tag{6}$$

où r et l dépendent des paramètres (μ, σ) selon (3).

2.2 Point fixe pour l'écart-type σ

Pour exprimer σ en fonction des quartiles des données tronquées, il est possible d'exploiter la forme paramétrique de la loi F_0 supposée $\mathcal{N}(\mu, \sigma)$. Soit Φ la fonction de répartition de la loi normale standard $\mathcal{N}(0, 1)$. On a donc $F_0(x) = \Phi\left((x - \mu)/\sigma\right)$ et il vient d'après (5) que pour $i = 1, 2, 3$

$$\frac{i}{4} = \frac{\Phi\left(\frac{q_{i,t} - \mu}{\sigma}\right) - \Phi(-\kappa)}{1 - 2\Phi(-\kappa)}.$$

Cette équation permet d'exprimer $q_{i,t}$ en fonction des paramètres μ et σ. En exploitant les symétries de Φ, i.e. $\Phi(-x) = 1 - \Phi(x)$, on peut ainsi exprimer l'écart interquartile des données tronquées en fonction de σ : $q_{3,t} - q_{1,t} = \sigma\,\lambda_\kappa$, où

$$\lambda_\kappa = 2\Phi^{-1}\left(\frac{1}{2}\left(\frac{1}{2} + \Phi(\kappa)\right)\right), \tag{7}$$

est une constante qui ne dépend pas des paramètres μ et σ, seulement du facteur κ de troncature (3). Ceci donne l'équation du point fixe suivante pour σ

$$\sigma = \frac{q_{3,t} - q_{1,t}}{\lambda_\kappa}, \tag{8}$$

où le second membre dépend des paramètres μ et σ à travers les expression des quartiles (4) et des bornes (3).

2.3 Algorithme du point fixe

La médiane μ et de l'écart-type σ sont définis comme les solutions des équations (6) et (8), i.e.

$$\mu = g_1(\mu, \sigma) = q_{2,t}, \tag{9}$$
$$\sigma = g_2(\mu, \sigma) = (q_{3,t} - q_{1,t})/\lambda_\kappa, \tag{10}$$

où les quartiles $q_{i,t}$ sont exprimés en (4) pour $i = 1, 2, 3$. Ces quantiles théoriques dépendent de la fonction de répartition F des données, qui est inconnue. Afin d'estimer les paramètres μ et σ, il est néanmoins possible de remplacer F par la fonction de répartition empirique des données $\bar{F}$, et F^{-1} peut alors être remplacée par l'inverse généralisée de $\bar{F}$. Ceci revient à remplacer les quantiles théoriques $q_{i,t}$ par leur estimateurs empiriques $\overline{q}_{i,t}$, comme détaillé dans l'algorithme 1.

Algorithme 1 FRONDE : σ-clipping par point fixe

Entrées : ensemble des données $I = \{x_1, \ldots, x_N\}$, facteur de troncature κ
$\mu_0 \leftarrow$ médiane empirique(I)
$\sigma_0 \leftarrow$ écart-type empirique(I)
$k \leftarrow 0$
while (μ_k, σ_k) n'a pas convergé **do**
  $k \leftarrow k + 1$
  $I_k \leftarrow \{x_i \ : \ |x_i - \mu_{k-1}| \leqslant \kappa\sigma_{k-1}\}$
  $\mu_k \leftarrow$ médiane(I_k)
  $\overline{q}_{1,t} \leftarrow$ quartile$(I_k, 25\%)$
  $\overline{q}_{3,t} \leftarrow$ quartile$(I_k, 75\%)$
  $\sigma_k \leftarrow (\overline{q}_{3,t} - \overline{q}_{1,t})/\lambda_\kappa$, où λ_k est défini en (7).
Sorties : μ_k, σ_k

Le facteur de troncature κ utilisé dans cet algorithme peut être fixé de manière à conserver un pourcentage p des données distribuées selon $\mathcal{H}_0$ autour de la médiane, i.e.

$$\kappa = \Phi^{-1}\left((1 + p)/2\right).$$

L'implémentation de l'algorithme 1 est disponible en Python [1] et en Matlab [2] sous le nom de FRONDE pour *Fixed-point algorithm for RObust Null Distribution Estimation*.

2.4 Convergence de l'algorithme

Supposons que le paramètre σ soit connu et fixé ($\sigma_k = \sigma$) dans l'algorithme 1 : seul le paramètre μ est mis à jour à chaque itération. Sans perte de généralité, on peut supposer que $\mu_1 \geq \mu_0$ (quitte à prendre l'opposé des observations). Il vient alors par récurrence que $\mu_{k+1} \geq \mu_k$ pour tout $k \geq 0$. Puisque par construction $\mu_k \leq \max_i x_i$, la suite μ_k est croissante et bornée, donc convergente. Par ailleurs, le nombre de données x_i est fini, donc le nombre de valeurs possibles pour les médianes μ_k l'est également. Ceci assure la convergence de l'algorithme en un nombre fini d'itérations.

1. https ://github.com/raphbacher/fronde-py
2. https ://github.com/raphbacher/fronde-matlab

Dans le cas général (σ inconnu), il est difficile d'étudier d'un point de vue théorique les conditions de convergence. Même si le nombre de valeurs possibles pour les paramètres μ_k et σ_k est par construction fini, il peut arriver en pratique que la suite des μ_k et σ_k décrive un cycle, en général d'ordre 1. Cependant un tel problème est rare dans nos simulations et peut aisément être détecté et corrigé : le paramètre σ est dans ce cas fixé à la valeur du cycle la plus conservative pour notre test d'hypothèses (dans l'optique de pouvoir toujours contrôler les erreurs de type I), i.e. la plus grande valeur σ_M du cycle. Seul le paramètre μ_k est ensuite mis à jour dans l'algorithme 1 où $\sigma_k = \sigma_M$, ce qui garantit la convergence d'après le paragraphe précédent.

2.5 Consistance des estimateurs

La fonction de répartition empirique $\bar{F}$ converge uniformément vers F lorsque le nombre N d'observations augmente (théorème de Glivenko-Cantelli). Par ailleurs, on peut démontrer que, sous P1, la fonction du point fixe

$$g(\mu, \sigma) = \left(g_1(\mu, \sigma), g_2(\mu, \sigma)\right),$$

définie par (9) et (10) à partir de la distribution F, est contractante au voisinage du point fixe. Par conséquent, les valeurs renvoyées par l'algorithme 1, qui reposent sur l'approximation empirique de $g(\mu, \sigma)$, convergent vers le point fixe de $g(\mu, \sigma)$. Ce point fixe correspond effectivement aux valeurs théoriques des paramètres de la loi F_0 sous P1, ce qui prouve la consistance des estimateurs.

3 Performances

La méthode proposée a été testée sur des données synthétiques similaires à celles de [5] afin de comparer les performances de notre méthode en toute impartialité. Pour chaque cas étudié, on génère 1000 cubes de taille $64 \times 64 \times 64$ pixels contenant un champ aléatoire gaussien généré par convolution d'un cube de données gaussienne $\mathcal{N}(0, 1)$ indépendantes avec un noyau 3D de forme gaussienne de moyenne nulle et d'écart-type 1.5 dans les 3 dimensions. Le champ gaussien a ensuite été décalé et mis à l'échelle afin de présenter une moyenne $\mu_0 = 0.2$ et un écart-type $\sigma_0 = 1.2$. Un signal constant d'amplitude 3 est ajouté dans un sous-cube de taille $T \times T \times T$ afin de simuler l'alternative $\mathcal{H}_1$. On fera varier la valeur de T afin d'étudier l'influence de la proportion $\pi_0 = 1 - T^3/64^3$ sur les performances d'estimation de μ_0 et de σ_0 (dans [5], $T = 16$).

On étudie des performances de l'algorithme 1 et de la méthode [3] présentée dans [4, 5] pour le seuillage de cartes statistiques en imagerie cérébrale IRM (détection d'activations, de connexions entre différentes zones du cerveau, etc). On compare également à la méthode dédiée décrite dans [3, chap. 6] [4] qui ajuste au sens du maximum de vraisemblance la distribution des données tronquées. Ces deux dernières méthodes

3. Une implémentation Python est disponible à l'adresse `http://nipy.sourceforge.net/nipy/devel/labs/enn.html` via la méthode `NormalEmpiricalNull.learn` du package `nipy`
4. Une implémentation R est disponible dans le paquetage **`locfdr`** à l'adresse `https://CRAN.R-project.org/package=locfdr`

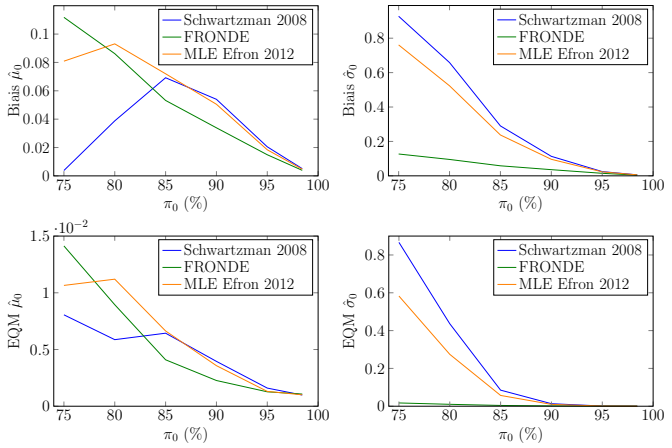

FIGURE 1 – Influence de la proportion π_0 sur le biais (ligne du haut) et l'EQM (ligne du bas) des estimateurs de moyenne μ (colonne de gauche) et d'écart-type σ (colonne de droite) pour les méthodes décrites dans [4] ("Schwartzman 2008"), [3] ("MLE Efron 2012"), et dans l'algorithme 1 ("FRONDE").

reposent sur une troncature préalable des données afin de ne conserver qu'une proportion p fixée des données les plus vraisemblables sous $\mathcal{H}_0$ pour faire l'estimation. A la différence de la méthode proposée dans l'algorithme 1, cette région de troncature n'est pas à mise à jour en fonction des paramètres estimés.

Les figures 1 et 2 illustrent les performances en terme de biais et d'erreur quadratique moyenne (EQM) des estimateurs $\hat{\mu}_0$ et $\hat{\sigma}_0$. Sur la figure 1 on fait varier π_0 de 75% (cas défavorable) à $98,44\%$ (cas plus facile). Pour les trois méthodes, $p = 80\%$ des données autour de la médiane sont conservées pour effectuer l'estimation. Pour $\pi_0 > 85\%$, nos estimateurs de moyenne et de variance sont sensiblement meilleurs que ceux de [4] et [3] que ce soit en terme de biais ou d'EQM. Inversement pour $\pi_0 < 85\%$, il semblerait que l'estimateur de moyenne soit moins biaisé pour [4] et [3]. Or si on décroit encore la valeur de π_0 le biais change de signe. En réalité c'est l'estimation très biaisée de la variance qui entraîne un décalage de la distribution sous $\mathcal{H}_0$ vers la gauche avec un passage fortuit par la bonne valeur de μ_0. Notre estimateur propose une estimation plus robuste de la variance.

Sur la figure 2 la proportion π_0 est fixée à 90% et on fait varier la proportion p de données conservées pour l'estimation entre 50% et 90%. Le choix de p reflète un compromis biais/variance : plus on garde de données plus on s'expose à un biais dus aux échantillons générés sous $\mathcal{H}_1$. Inversement, il faut conserver une certaine proportion de données pour garantir une variance des estimateurs raisonnable. La figure 2 illustre les performances de la méthode proposée en terme de biais et d'EQM par rapports aux deux autres méthodes, en particulier lorsque $p > 70\%$. La précision de l'estimation de l'hypothèse $\mathcal{H}_0$ est capitale lors du seuillage des données par la procédure de contrôle du FDR de Benjamini-Hochberg [7]. La figure 3 montre ainsi les performances de détection obtenues en fonction de la proportion π_0, de la proportion p de données conservées et des méthodes d'estimation. Ces résultats confirment ceux présentés dans les figures 1 et 2. Le gain en puissance obtenu pour la méthode proposée (tout en conservant le contrôle) y apparaît en effet clairement. Ceci illustre l'intérêt de la méthode proposée, notamment par rapport à ce qui est utilisé actuellement dans le paquetage `nipy` pour le traitement de données IRM, et plus généralement pour des applications de tests multiples.

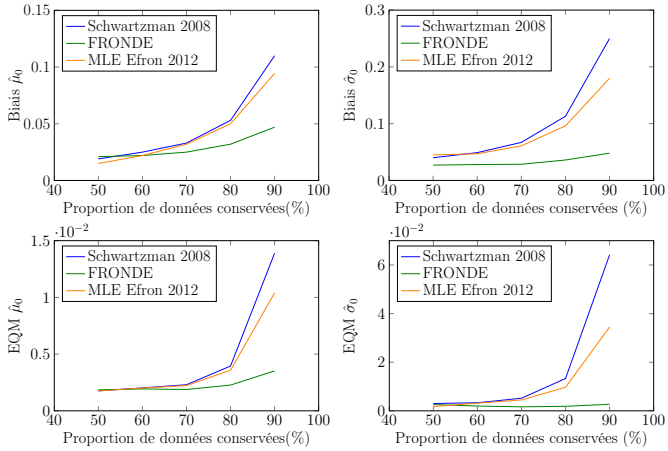

FIGURE 2 – Influence de la proportion p de données conservées dans la troncature sur le biais (ligne du haut) et l'EQM (ligne du bas) des estimateurs de moyenne μ (colonne de gauche) et d'écart-type σ (colonne de droite).

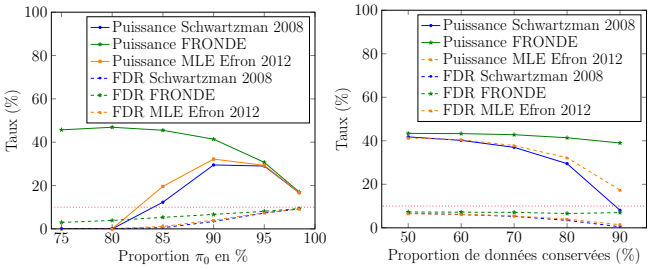

FIGURE 3 – Influence de la proportion π_0 (gauche) et de la proportion p de données conservées dans la troncature (droite) sur les puissances (traits pleins) et le FDR (tirets) obtenus pour un contrôle nominal du FDR à 10% (ligne horizontale en pointillés rouge) par la procédure de Benjamini-Hochberg selon les différents estimateurs des paramètres μ et σ.

Références

[1] D. Donoho et al. *Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects*. Statistical Science, 2015.

[2] C. Meillier et al. *Error control for the detection of rare and weak signatures in massive data*. EUSIPCO, 2015.

[3] B. Efron. *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction*.Cambridge Univ. Press, 2012.

[4] A. Schwartzman et al. *Empirical null and false discovery rate inference for exponential families*, The Annals of Applied Statistics, 2008.

[5] A. Schwartzman et al. *Empirical null and false discovery rate analysis in neuroimaging*. NeuroImage, 2009.

[6] R. Maronna, et al. *Robust statistics*, John Wiley and Sons, 2006.

[7] Y. Benjamini et al. *Controlling the false discovery rate : a practical and powerful approach to multiple testing*. Journal of the royal statistical society, 1995.

[8] R. Bacher et al. *Robust control of varying weak hyperspectral target detection with sparse non-negative representation*. Arxiv preprint. To appear in IEEE TSP, 2017.