Global error control procedure for spatially structured targets

Raphael Bacher*†,
Florent Chatelain* and Olivier Michel*
*University Grenoble-Alpes, Gipsa-Lab, Grenoble, France
†CRAL, Observatoire de Lyon, Lyon, France
Email: raphael.bacher@gipsa-lab.fr

Abstract—**In this paper, a target detection procedure with global error control is proposed. The novelty of this approach consists in taking into account spatial structures of the target while ensuring proper error control over pixelwise errors. A generic framework is discussed and a method based on this framework is implemented. Results on simulated data show conclusive gains in detection power for a nominal control level. The method is also applied on real data produced by the astronomical instrument MUSE.**

I. Introduction

In this paper we consider the problem of detecting spatial extension of possible targets/sources in massive datasets based on a proper statistical control of error rates. In this context of multiple testing, the classical type I error control of each individual test might not be appropriate; see e.g., [1], [2] for applications in image processing. The number of wrongly rejected null hypotheses can become important (i.e., even larger than the number of true detections) due to the high number of pixels to be tested. To address this issue, a global error control approach, namely the False Discovery Rate (FDR), was introduced in [3].

A simple and widely used approach to control FDR is the Benjamini and Hochberg (BH) procedure developed in [3]. This latter only requires knowledge of the test statistics distribution for noise-only observations. However, for real-word data, this distribution is unlikely to be known exactly. For instance, when the noise exhibits a complex, possibly heteroscedastic dependence structure, a robust estimator for the distribution from noise-plus-target observations may be difficult to obtain. In addition, applying BH procedure in a pixelwise framework does not take into account any relation between targeted samples. In many applications targeted samples exhibit some coherence, for instance, pixels of interest are often organized in connected structures. Taking this kind of prior into account can only improve detection power. Note that some groupwise approaches were developed recently [4] but require an *a priori* knowledge of the groups, in general not available as the target shape is *a priori* unknown. Lastly, a drawback of global error control is that detection power may decrease as the number of tests increases. For instance, the detection power of a single target, at a given nominal FDR, depends of the size of the region into which it is sought.

Funding by ERC grant 339659-MUSICOS

Thus, in order to 1) ensure a robust error control, 2) account for target structure and 3) mitigate the influence of the number of tests, a new simple framework is proposed called COnnection accounting Method for Extracting Target (COMET). COMET elaborates on both a recent alternative class of FDR controlling procedures [5], [6] and a recent method [7] that yields robust control under weak assumptions on the noise. We propose a relevant framework in the context of multiple-testing problem where target exhibit connected structures. The control is based on some feature statistics built from the data that are designed to satisfy some symmetry properties. This context covers a large number of applications. Hereinafter we focus on detection problems where noise is assumed symmetrically distributed and target have a positive contribution. Detection of galactic halo in hyperspectral data satisfies these assumptions. Using a simple and robust procedure [7], such feature statistics are built and motivated the method presented in this paper.

Section II presents the proposed method for detecting a multi-pixels target with strong connectivity while ensuring global control. Section III illustrates the performance of COMET on simulated data. In section IV this method is tested on real data produced by MUSE instrument [8]. Finally some conclusions and perspectives are drawn in section V.

Notations

In our context we use interchangeably the words "pixel" or "sample". Data vector associated with a pixel (a spectrum in hyperspectral context) is represented by bold letters e.g. $\boldsymbol{y}$. $a \vee b$ denotes $\max(a, b)$.

II. Method

We address here the detection of a positive signal, from noisy sample vector $\boldsymbol{y} \in \mathbb{R}^l$. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be the hypotheses denoting respectively the absence or presence of the source contribution. Thus we get the following one-sided testing problem:

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{y} = \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \boldsymbol{y} = \alpha \boldsymbol{d} + \boldsymbol{\epsilon}, \quad \text{with } \alpha > 0, \end{cases} \tag{1}$$

$\boldsymbol{\epsilon} \in \mathbb{R}^l$ is a noise vector with unknown symmetrical distribution. $\boldsymbol{d}$ is a known template or reference vector. $\boldsymbol{y} \in \mathbb{R}^l$ may be a spectrum in hyperspectral imaging, a time series in

ISBN 978-0-9928626-7-1 © EURASIP 2017

fMRI imaging, or an intensity ($l = 1$ and $\boldsymbol{d} = 1 \in \mathbb{R}$) in conventional imaging.

We now address the problem of solving (1) for a large number n of data realizations $\{\boldsymbol{y}_i\}_{1 \leq i \leq n}$. As stressed earlier, in the context of multiple testing, a classical type I error control of each individual test might not be appropriate. Here our purpose is to compute a data-dependent decision threshold controlling the FDR introduced in [3]. FDR is the expected proportion of true null hypotheses wrongly rejected, which are referred to as the *false discoveries*, among all of the rejected tests:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{U}{R \vee 1}\right],$$

where R is the total number of tests for which the null hypothesis is rejected, while U is the number of false discoveries among the R discoveries. A new method for controlling FDR was recently proposed by Barber and Candès in [5] based on the construction of knockoffs. These knockoffs allow them to build feature statistics $\{w_i\}_{1 \leq i \leq n}$ for each test which are, in particular,

- symmetrical under $\mathcal{H}_0$, i.e. $\mathbb{P}(w_i > 0 | i \in \mathcal{H}_0) = \mathbb{P}(w_i < 0 | i \in \mathcal{H}_0)$,
- stochastically larger under $\mathcal{H}_1$ than under $\mathcal{H}_0$, i.e. $\mathbb{P}(w_i > 0 | i \in \mathcal{H}_1) > \mathbb{P}(w_i > 0 | i \in \mathcal{H}_0)$.

Barber and Candès then established some FDR control procedures that we reformulate here in our context.

Proposition II.1. *If the $\{w_j\}_{1 \leq j \leq n}$ are symmetrically distributed under $\mathcal{H}_0$, and their signs are independent, then for a nominal control level q, thresholding at level*

$$\widehat{t_q} = \inf\left\{t \geq 0 : \frac{1 + \#\{w_j < -t\}}{1 \vee \#\{w_j > t\}} \leq q\right\} \tag{2}$$

ensures an exact control of the FDR at level q for the set of detections $\mathcal{D} = \{i : w_i > \widehat{t_q}\}$.

Proof. See Theorem 3 of [5], applied for binary p-values

$$p_i = \begin{cases} 1/2, & \text{if } w_{(i)} > 0, \\ 1, & \text{if } w_{(i)} < 0, \end{cases}$$

where $w_{(i)}$ are the absolute ordered feature statistics such that $|w_{(1)}| \geq |w_{(2)}| \geq \cdots \geq |w_{(n)}|$. □

A. Building of feature statistics

In [5], feature statistics are based on knockoffs, that are artificial variables that mimic the correlation structure of the original ones, in a context of linear regression problems with white Gaussian noise. This construction may be challenging in practice, especially in high dimension. We propose here a simpler feature construction, adapted to the detection problem exposed in (1), where noise is symmetrically distributed and target has a positive ($\alpha > 0$) contribution. For known $\boldsymbol{d}$ the feature w_i for the ith test, $1 \leq i \leq n$, can be defined as:

$$w_i = \boldsymbol{d}^T \boldsymbol{y}_i. \tag{3}$$

which is the matched filter statistics for white noise. Thus a high positive value of w_i indicates most certainly a non-null sample and under $\mathcal{H}_0$, the w_i are symmetrically distributed (as noise is assumed to be symmetrically distributed).

We now consider a multiple-testing detection problem where target samples are expected to be structured (a natural example is a multi-pixel object in an image). We propose a generic framework accounting for spatial connexity to improve the detection power of the test (2) while ensuring the same FDR control.

B. Generic framework

The main idea is the following: as target is expected to form connected areas, only pixels in the neighborhood of already detected pixels are tested. This amounts to develop a region growing approach. Reducing the number of tests to the neighboring pixels allows to limit the loss in power for a given FDR level. This is performed by the step-up procedure described in Algorithm 1. At a given selection step, let $\mathcal{A} \subset \{1, \ldots, n\}$ be the set of "interesting" pixels[1] obtained for a given selection step. Ensuring FDR control with this Algorithm 1 procedure is then equivalent to guaranty P1 for the selection procedure:

P1 (Post-selection symmetry)**.** *For any pixel $j \in \mathcal{A}$ corresponding to a true null hypothesis, w_j is symmetrically distributed.*

P1 means that the selection step preserves the symmetry under $\mathcal{H}_0$. At iteration k, let $\mathcal{A}_k$ be the set of selected pixels and $\mathcal{D}$ the final set of detected pixels at the end of the procedure. A false discovery proportion (FDP) estimate (among the positive features $w_i > 0$ selected in $\mathcal{A}_k$) is defined as

$$\widehat{q_k} = \frac{1 + \#\{i \in \mathcal{A}_k, w_i < 0\}}{1 \vee \#\{i \in \mathcal{A}_k, w_i > 0\}}. \tag{4}$$

Finally S denotes the operator associated with the selection procedure. This operator promotes spatial connectivity in newly selected pixels. The following lemma can easily be established due to the noise distribution symmetry.

Lemma II.2. *If the selection operator S depends on the data only through the absolute values $|w_i|$ of the feature statistics, for $1 \leq i \leq n$, then property P1 is satisfied.*

Note that this framework allows for numerous implementations of the selection procedure. This consists mainly of finding a good (promoting true $\mathcal{H}_1$ samples) ordering of the pixels to be tested. In the following we focus on a simple selection procedure that gives satisfactory results in practice.

C. Proposed implementation

The following greedy approach is proposed: at each step the greatest feature in absolute value is retained among the neighboring pixels. At step k, let $\mathcal{N}_k = G(\mathcal{A}_k)$ be the external neighborhood of $\mathcal{A}_k$, where G is the morphological external gradient, i.e. a dilation (here for a 8-connectivity clique)

[1] For the sake of simplicity, each one of the n pixels is identified by an index $i \in \{1, \ldots, n\}$ rather than by its spatial coordinates. Note however that, hereinafter, connectivity has to be understood w.r.t. spatial coordinates.

Algorithm 1 Generic COMET step-up procedure

Inputs: feature statistics $\boldsymbol{w} = \{w_j\}_{1 \le j \le n}$, nominal control level q

$k \leftarrow 0$, $\mathcal{A}_0 \leftarrow \emptyset$, $\widehat{q}_0 \leftarrow 0$ ▷ loop initialization
while $\mathcal{A}_k \neq \{1, \dots, n\}$ **do**
  $\mathcal{A}_{k+1} \leftarrow S(\mathcal{A}_k, \boldsymbol{w})$ ▷ selection step satisfying P1
  compute $\widehat{q}_{k+1}$ using (4) ▷ FDP estimate update
  $k \leftarrow k + 1$
endWhile
$\widehat{k} \leftarrow \max\{k : \widehat{q}_k \le q\}$ ▷ step-up stopping rule
Output: $\mathcal{D} \leftarrow \{i \in \mathcal{A}_{\widehat{k}} : w_i > 0\}$ ▷ list of detections

followed by a subtraction. Then the selection procedure is defined as

$$S(\mathcal{A}_k, \boldsymbol{w}) \equiv \mathcal{A}_k \cup \{j_0\}, \quad \text{where } j_0 = \arg\max_{j \in \mathcal{N}_k} |w_j|.$$

The symmetry property P1 is ensured by lemma II.2. In practice, to improve computation times, the inner loop of the step-up procedure can be stopped when both the number of selected pixels is large and $\widehat{q}_k$ is significantly greater than q (e.g. $\widehat{q}_k \ge 1.2 \times q$).

D. Control under independent noise

Proposition II.3 (FDR control of COMET)**.** *Assume that the noise vectors* $\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_n$ *are symmetrically distributed and independent. Then Algorithm 1, where the feature statistics* $\boldsymbol{w}$ *are built using* (3)*, ensures an exact control of the FDR:* $\mathbb{E}\left[\frac{U}{R \vee 1}\right] \le q$.

Proof. According to property P1, for all $i \in \mathcal{A}_{\widehat{k}}$ corresponding to a true $\mathcal{H}_0$, w_i is symmetrically distributed. Moreover the signs of $\{w_i\}_{1 \le i \le n}$ are independent as are the $\{\boldsymbol{\epsilon}_i\}_{1 \le i \le n}$. Thus we can apply proposition II.1 where the threshold reduces to $\widehat{t}_q = 0$. This concludes the proof. □

We stress that this is quite a strong result as neither stationarity nor full knowledge (i.e. aside its symmetry property) of the noise distribution are required.

E. Control under correlated noise

In presence of noise correlations, independence of the $\{w_i\}$ signs is in general not satisfied. Consequently we can no longer rely on proposition II.3. Nevertheless we observe that an asymptotic control is attained. This can be proved using the following assumptions.

A1 (Weak dependence)**.** *For any set of* $\mathcal{H}_0$ *pixels* $\mathcal{S}$,

$$\frac{\#\{i \in \mathcal{S} : w_i > 0\}}{\#\mathcal{S}} \underset{\#\mathcal{S} \to \infty}{\overset{a.s.}{\longrightarrow}} \mathbb{P}(w_i > 0)$$

To simplify notation, $\mathcal{S}_n$ now denotes the final selection set $\mathcal{A}_{\hat{k}}$ given by Algorithm 1 for a size n set of pixels.

A2 (Region growth)**.** *For a given* $q > 0$, $\mathcal{S}_n$ *increases with the number of samples* n *and* $\#\mathcal{S}_n \underset{n \to +\infty}{\longrightarrow} +\infty$

TABLE I: Comparison of power between COMET and a similar procedure without taking connectivity into account. Noise is Gaussian correlated. Nominal FDR is 5%. Target is 250 pixels wide, total number of pixels is 2601 (51×51) for the first two areas and 71×71 for the larger area. Results are averaged on 400 Monte-Carlo runs.

	COMET	Non-connected procedure
Source + noise		
False discovery rate (%)	4.85	4.74
Power (%)	76.1	35.4
Source + noise on larger area		
False discovery rate (%)	4.84	4.65
Power (%)	76.0	25.5

Proposition II.4 (Asymptotic control of COMET)**.** *Assume A1, A2 and symmetrical distribution of feature statistics under* $\mathcal{H}_0$*. Then Algorithm 1 ensures an asymptotic control of the FDR of the test* (1).

Sketch of the proof. For a given control q and a set of n samples, let $\widehat{\text{FDP}}_q = \frac{1 + \#\{i \in \mathcal{S}_n, w_i < 0\}}{\#\{i \in \mathcal{S}_n, w_i > 0\}}$. Based on the above assumptions, it can be shown that $\liminf_{n \to \infty}(\widehat{\text{FDP}}_q - \text{FDP}) \ge 0$. Since $\widehat{\text{FDP}}_q \le q$ due to the stopping rule, it comes from Fatou's lemma that $\limsup_{n \to \infty} \mathbb{E}[\text{FDP}] = \text{FDR} \le q$ □

F. Generalization to sparse non-negative representation

COMET can be exploited in the framework of sparse non-negative representation such as developed in [7]. Each feature is now defined as

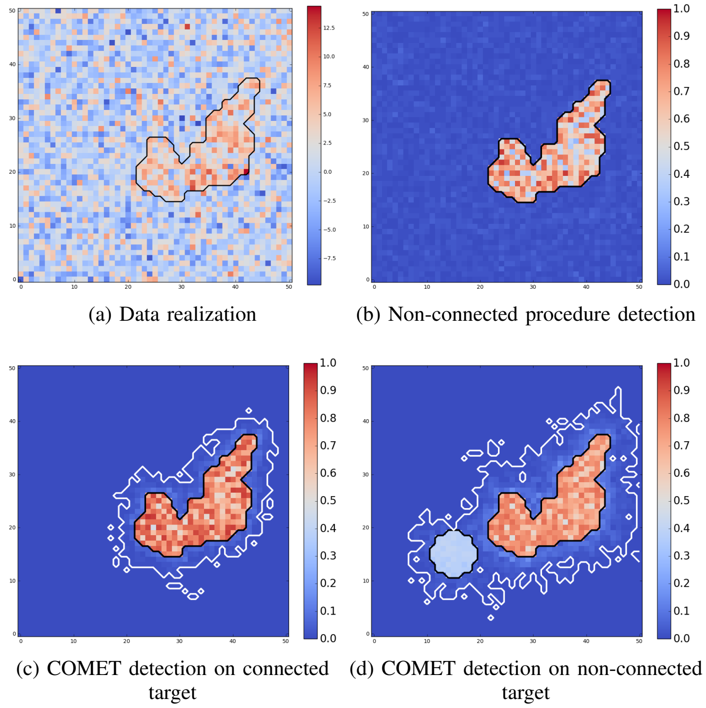
$$w_i = w_i^+ \vee w_i^- \times \begin{cases} +1 \text{ if } w_i^+ > w_i^-, \\ -1 \text{ if } w_i^+ < w_i^-, \end{cases} \tag{5}$$

where for $1 \le i \le n$, $w_i^+ = \max_j \{\boldsymbol{d}_j^T \boldsymbol{y}_i\}$ and $w_i^- = -\min_j \{\boldsymbol{d}_j^T \boldsymbol{y}_i\}$. $\boldsymbol{D} = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_k]$ is a dictionary of templates designed to account for possible variability of the target. For a symmetrically distributed noise, these features are also symmetrical under $\mathcal{H}_0$ and high positive values are expected for target pixels (see [7]).

III. Simulation and Results

In this section, COMET procedure is compared with the method described in proposition II.1, that does not take connectivity into account. For sake of illustration, we consider datacubes of dimension $31 \times 51 \times 51$ that mimic hyperspectral data (one spectral dimension, 2 spatial dimensions). Additive Gaussian noise is convolved by a 3×3 kernel to introduce spatial correlations. A target is composed of 250 connected pixels with associated spectra. Each spectrum is a truncated (± 6 spectral bands around the mode) Gaussian shaped density (with $\sigma \approx 2.1$) centered on the median band $j = 15$. $\boldsymbol{D}$ contains 3 spectral templates obtained by shifting the mode by ± 1 spectral band. Data peak signal-to-noise ratio pSNR is around 5dB (pSNR is defined as $\text{pSNR} = 20 \log \frac{u_m}{\sigma}$ where u_m is the maximal value of target data and σ is the noise standard deviation).

As shown in table I, taking connectivity into account drastically improves the detection power while preserving FDR control. Let us emphasize that both FDR procedures adapt to the data (contrary to a control based on PFA level): in the absence of target, detection almost never occurs, thus ensuring the control of the expected FDP (note that in the case of occurrence of a false detection, the FDP rises to 100%). Moreover our results illustrate the robustness of COMET against the size of the region to be tested: detection power remains approximately constant. Figure 1 illustrates the detection procedure: figure 1(a) represents a noisy realization of tested data, summed over the spectral channels ("white image"); figures 1(b) and 1(c) show the detection rate over 100 Monte-Carlo runs for respectively the non-connected approach and COMET. Color coding emphasizes the gain of power obtained with COMET. The region explored by COMET (white contour) stays in the vicinity of the target; on the contrary the method without connectivity prior leads to largely spread false discoveries (almost all pixels are detected at least once in fig. 1b), due to spurious noise peaks. Figure 1(d) illustrates the performance of COMET when the connectivity assumption is violated: a disconnected "blob" can still be detected and overall detection power is still in favor of COMET (around 66,4% vs 54.5% without connectivity prior for shown example). Figure 2 depicts the procedure performances for several FDR level: figure 2(a) underlines that asymptotic control is indeed obtained in presence of correlated noise, for both approaches; figure 2(b) again emphasizes the significant gain in power due to taking connectivity into account.



(a) Data realization

(b) Non-connected procedure detection

(c) COMET detection on connected target

(d) COMET detection on non-connected target

Fig. 1: (a) example of a noisy datacube realization (summed over the spectral dimension); (b) map of detection rate for an 10% FDR procedure without connectivity prior (1 means always detected, 0 means never detected); (c) map of detection rate at same 10% FDR level for COMET; (d) map of detection rate for COMET on non-connected target. Results averaged on 100 Monte-Carlo runs of simulated data. Black line indicates ground-truth target, white line indicates pixels detected at least once among all Monte-Carlo runs with COMET.

IV. Application to MUSE data

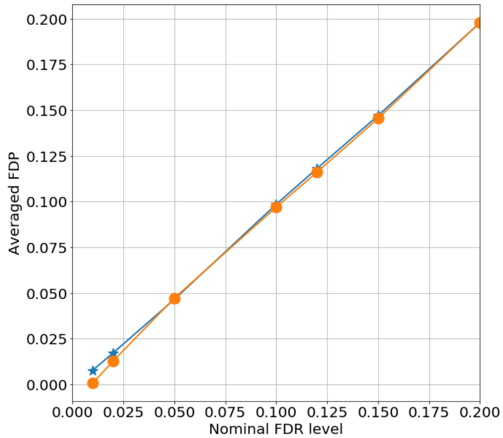
In [7], a detection method for galactic halos in hyperspectral data produced by the MUSE instrument [8], [9] was developed. This method was based on BH procedure which did not take connectivity into account, whereas halos are expected to exhibit strong connectivity. Using the feature construction described in II-F, the COMET procedure can be applied for this halo detection problem. Initialization of the detection process (design of the dictionary, choice of the region to explore) benefits from the pre-detection of the galactic core by specifically designed method such as [10].

Figure 3 illustrates the results obtained on an object of the Hubble Deep Field South (HDFS) MUSE observation [9]: figure 3(a) depicts the feature statistics map of the explored region; figure 3(b) and 3(c) show the detection map for a 10% FDR level for respectively the non-connected approach and COMET; figure 3(d) underlines the difference between the two detection maps. Adding connectivity prior for halo detection in MUSE data increases detection power: twice more pixels are detected for the studied object. The precise gain in detection power is difficult to evaluate as there is of course no ground truth available. As expected, COMET favors connected pixels detection so sometimes isolated pixels are detected only by the classical approach. Nonetheless halos are expected to be connected so these tiny "blobs" are likely to be spurious detections. It is thus obvious that COMET improves here detection power while preserving from spurious noise peaks.

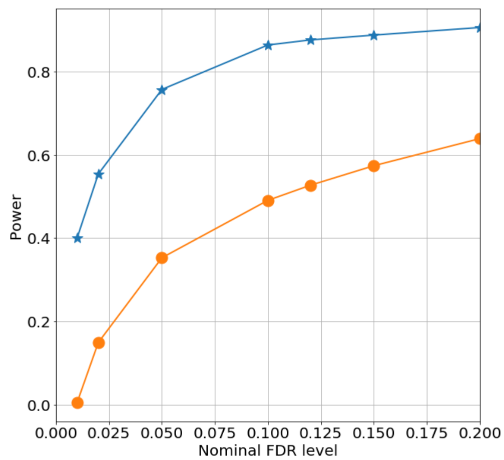
V. Conclusion

In this paper a generic framework for improving detection power of FDR methods on connected multi-pixels targets is proposed. This framework is based on selecting pixels to be tested using connectivity constraints. An implementation of a simple procedure for this framework, named COMET, is described; it only requires symmetry of the noise distribution and positivity of the targets. Exact FDR control of the procedure is proven for independent samples. When noise is weakly dependent, asymptotic FDR control is proven. Compared to a state of the art FDR procedure without connectivity prior, COMET exhibits conclusive results on simulated data: we obtain both a substantive gain in detection power for a same nominal FDR level and a good robustness to the size of the region to be tested. This procedure is applied on real data from the MUSE instrument with promising results. This method is non-parametric, thus robust to model mis-specification. Finally

ISBN 978-0-9928626-7-1 © EURASIP 2017

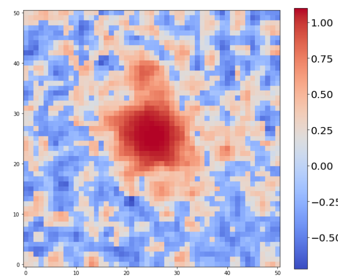


(a) FDR control

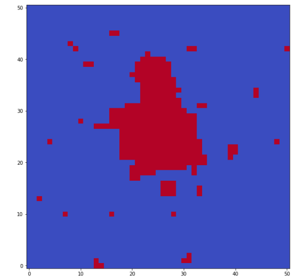


(b) Detection power

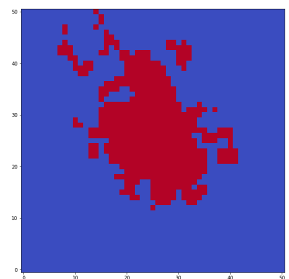
Fig. 2: (a) Empirical FDR vs nominal FDR ; (b): power vs nominal FDR. Proposed method is in blue ($*$), non-connected approach is in orange (●). Results averaged on 2000 Monte-Carlo runs of simulated data, with a Gaussian noise correlated by a 3×3 kernel. pSNR is 5dB.



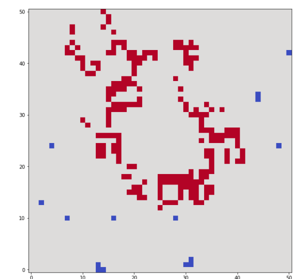
(a) Feature statistics map



(b) Detection map of non connected approach



(c) Detection map of COMET



(d) Difference between detection maps

Fig. 3: Detection maps for object ID92 of the HDFS MUSE observation [9] at a nominal FDR of 10%. (a) feature statistics map of considered region; (b) detection map of the classical method without connectivity prior; (c) detection map produced by COMET; (d) difference between the two map. Detected pixels are in red in (b) and (c). In (d), pixels detected only by COMET are in red, pixels only detected without connectivity prior are in blue.

it has a low computational cost (around 1s to process a $31 \times 51 \times 51$ datacube).

Acknowledgment

The authors would like to thank the ERC grant 339659-MUSICOS for the funding of this work and Roland Bacon for his astrophysical expertise and the access to MUSE data.

References

[1] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *Neuroimage*, vol. 15, no. 4, pp. 870–878, 2002.

[2] C. Meillier, F. Chatelain, O. Michel, and H. Ayasso, "Error control for the detection of rare and weak signatures in massive data," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1974–1978.

[3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[4] R. F. Barber and A. Ramdas, "The p-filter: multi-layer fdr control for grouped hypotheses," *arXiv preprint arXiv:1512.03397*, 2015.

[5] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, pp. 2055–2085, 2015.

[6] E. Candès, Y. Fan, L. Janson, and J. Lv, "Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection," *arXiv preprint arXiv:1610.02351*, 2016.

[7] R. Bacher, C. Meillier, F. Chatelain, and O. Michel, "Robust control of varying weak hyperspectral target detection with sparse nonnegative representation," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3538–3550, 2017.

[8] R. Bacon, M. Accardo, L. Adjali, H. Anwand, S. Bauer, I. Biswas, J. Blaizot, D. Boudon, S. Brau-Nogue, J. Brinchmann *et al.*, "The muse second-generation vlt instrument," in *SPIE Astronomical Telescopes+ Instrumentation*. International Society for Optics and Photonics, 2010, pp. 773 508–773 508.

[9] R. Bacon, J. Brinchmann, J. Richard, T. Contini, A. Drake, M. Franx, S. Tacchella, J. Vernet, L. Wisotzki, J. Blaizot *et al.*, "The MUSE 3D view of the Hubble Deep Field South," *A&A*, vol. 575, p. A75, 2015.

[10] C. Meillier, F. Chatelain, O. J. Michel, and H. Ayasso, "Nonparametric bayesian extraction of object configurations in massive data." *IEEE Trans. Signal Processing*, vol. 63, no. 8, pp. 1911–1924, 2015.