ERROR CONTROL FOR THE DETECTION OF RARE AND WEAK SIGNATURES IN MASSIVE DATA

Céline Meillier, Florent Chatelain, Olivier Michel, Hacheme Ayasso

GIPSA-lab, Grenoble Alpes University, France

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 are derived for independent tests. 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) control can be directly applied to the result produced by a very common tool in signal and image processing: the matched filter. This shows that despite the dependency structure between the components of the matched filter output, the Benjamini-Hochberg procedure still guarantee the FDR control. This is illustrated on both synthetic and real data.

***Index Terms*—** source detection, matched filter, error control, FDR, massive data

1. INTRODUCTION

In this work we consider the case of noisy observation data of some sources whose responses are faint and sparse in the observation domain. The normalized signatures of these sources are supposed to be known but their position and their intensity must be estimated. This is the case, for instance, of images containing unresolved sources whose response is the point spread function (PSF) of the optical device. We suppose that these sources contribute to a low proportion of data samples (i.e. the pixels in the image case), while the other samples are just corrupted by Gaussian noise. We want to detect the largest number of samples belonging to a source while controlling the number of *false discoveries*, i.e. samples that contain only noise but were detected as a source contribution.

This leads to consider the following statistical linear observation model:

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

where $Y \in \mathbb{R}^d$ is the observation vector (for an image, Y correspond to the vectorized image), $\epsilon \in \mathbb{R}^d$ is a white noise vector (measurement, environment, etc), $H \in \mathbb{R}^{d \times n}$ is a design matrix. It may be a dictionary of possible sources and $\boldsymbol{a} \in \mathbb{R}^n$ the intensity vector. A few coefficients a_i of $\boldsymbol{a}$ are expected to be non zero, and correspond to the intensities of the few sources that are observed.

Model selection procedures are widely used to tackle this detection problem. They are commonly designed to minimize a penalized least-squares criterion $\widehat{\boldsymbol{a}} = \arg\min_\alpha ||Y - H\boldsymbol{a}||_2^2 + \lambda \text{Pen}(\boldsymbol{a})$, where the penalization $\text{Pen}(\boldsymbol{a})$ (e.g. ℓ_1 norm, ℓ_0 pseudo-norm or AIC criterion, etc) is chosen to enforce sparsity. This can be viewed as a soft, or hard, thresholding of the least squares estimator of the intensity vector. However a global error control cannot be guaranteed. This control is of particular interest for the detection of rare and weak sources, since this provides an interpretable criterion to achieve the tradeoff between the detection power and the number of false detections.

Deciding which samples may belong to a source can be formulated as a multiple hypotheses test on the intensity vector coefficients:

$$\begin{cases} \mathcal{H}_0^{(i)} & : \quad a_i = 0 \qquad \text{(noise only)} \\ \mathcal{H}_1^{(i)} & : \quad a_i > 0 \quad \text{(source + noise)} \end{cases}$$

Thresholding the intensity vector is then equivalent to reject the $\mathcal{H}_0^{(1)}, \dots, \mathcal{H}_0^{(n)}$ hypotheses for a given significance level or error criterion. When the threshold value is set to correspond to a significance level α valid for a single test, it does not take into account the large number of tests. This may result in situations where most of the detections over the n tests correspond to false alarms. Applying a Bonferonni correction [2] for the n tests yields a significance level of α/n which ensures to control the probability α of making even one false alarm. However this naive procedure is highly conservative and most of the true $\mathcal{H}_1^{(i)}$ are missed. This underlines that simple thresholding procedures are not efficient when a large number of tests is performed. This is why some authors have focused on more powerful procedures to monitor a global error rate. A very popular and attractive procedure, introduced by Benjamini and Hochberg [3], is to control the false discovery rate (FDR), *i.e.* the expected proportion of true null hypotheses rejected among all the rejected tests. The FDR control procedures have been recently studied in many different fields: astronomy [4,5], in functional neuroimaging [6] or in genomics [7]. Note that in [8] the authors propose a variable selection procedure, called the knockoff filter, designed

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

to control the FDR criterion in the statistical linear model (1) that we are considering. However in the case of massive data, both the prohibitive number of tests ($n \sim 10^8$ in the proposed application of this paper) and the strong local correlations between the test statistics make the building of knockoffs impossible.

In our case, the sources are very faint. Since their signature is assumed to be known, the matched filter is a classical approach to increase their detectability. This strategy is used in many application fields: in biology [9] for improving the segmentation of blood vessels from the background of retina images; in astronomy, for detection of point sources in cosmic microwave background images [10]. However, matched filtering introduces correlations between the output components and consequently between tests. Most of the FDR control procedures are designed in the case of independent tests. Note that a log-factor correction to the Benjamini-Hochberg (BH) procedure is proposed in [1]. This allows to control the FDR criterion under any arbitrary dependency structure. Unfortunately, this correction appears to be too conservative to be useful in practice.

Finally, detection of rare and weak signals in massive data issue can be split in two steps: 1) finding a list of samples/pixels that probably belong to sources with a global error control and then 2) identifying sources based on these discoveries with dedicated algorithms. In this paper, we focus on the first step to perform the screening of samples that may be affected by a source. In [1] the authors also show that under some assumption on the dependency structure, the BH procedure can still be applied. Based on this result, we show that under classical assumptions on the positivity of the source responses, the BH procedure for FDR control can be directly extended to the matched filter output.

2. PROBLEM FORMULATION

2.1. Model

The source signatures are assumed to be sparse and their profile is convolved with the response of the measuring instrument. The observation data can be decomposed into sources contribution and additive Gaussian noise:

$$Y = \sum_j s_j + \epsilon$$

where s_j is the response of the instrument to the j^{th} source and ϵ is the background noise. The source response can be modeled as:

$$s_j(r) = \sum_i a_{j,i} h(r - r_i)$$

where $a_{j,i}$ is the j^{th} source intensity at position r_i, i ranges over the source support, r is a coordinate in the data domain and h is the ℓ_2-normalized profile of the instrument response. Sources are assumed to be sparse, thus their support is assumed to be small. Note that data can be multidimensional (for an image, the coordinate r represents the cartesian coordinates (x, y)).

2.2. Assumptions

The major assumptions concern the positivity of the sources response: $a_{j,i} > 0$ for all i ranging over the j^{th} source support and for all sources j, and the response of the instrument is non-negative: $\forall r, h(r) \geqslant 0$. Note that this last assumption can be relaxed. If h contains a few negative coefficients, a sub-optimal matched filter will be applied based on the positive truncated template $h^+(r) = h(r)$ if $h(r) > 0$, $h^+(r) = 0$ otherwise, which yields template non-negativity.

We also assume that the additive noise vector ϵ is Gaussian distributed: $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ where σ^2 is known. Without loss of generality, let $\sigma^2 = 1$.

2.3. Detection strategy

As the number of sources and their positions are unknown, we have to test the d possible locations and we have to estimate the coefficients of the vector $\boldsymbol{a}$. In our case, the source detection strategy consists on using the linear model (1), where H is the $n \times n$ matrix (in this case $n = d$). H is defined columnwise. Each column is a shifted version of the ℓ_2-normalized response h, center at position i, $1 \leqslant i \leqslant n$. The intensity coefficients $a_i = \sum_j a_{j,i}$, obtained by summing all the source intensities at a given position r_i, are stored in the $1 \times n$ vector $\boldsymbol{a} = [a_1, \cdots, a_n]^T$. This leads to test the presence of a source at each position r_i:

- if $a_i > 0$ there is a source contribution at position r_i,
- if $a_i = 0$ there is no source at this location.

2.4. Matched filter and positive covariances

In order to increase the sources SNR, the optimal matched filter is defined as the inner product with the response h to be detected. In the matrix formulation, applying this is performed by the following operation:

$$H^T Y = H^T H \boldsymbol{a} + H^T \epsilon \qquad (2)$$

Note that matrix H has interesting properties:

P1. H is sparse, the number of non null coefficients in each column is small compared to d.

P2. $(H^T H)_{i,i} = 1 \; \forall \; 1 \leqslant i \leqslant n.$

P3. H is a non-negative matrix: $H \geqslant 0$ (*i.e.* all the matrix entries are non-negative) since the profile h is positive.

From (2), the $H^T Y$ vector is Gaussian distributed:

$$H^T Y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (3)$$

where:

- $\boldsymbol{\mu} = H^T H \boldsymbol{a} \geqslant 0$,
- $\boldsymbol{\Sigma} = H^T H \geqslant 0$.

3. MULTIPLE HYPOTHESES TESTING AND FDR CONTROL

3.1. Multiple hypotheses testing formulation

Considering the matched filter output $H^TY = X$ for reading simplicity, from prop. P2 and (3), each component X_i is Gaussian distributed:

$$X_i \sim \mathcal{N}(\mu_i, 1), \; \forall\, 1 \leqslant i \leqslant n$$

We now consider the following binary test for each X_i:

$$\begin{cases} \mathcal{H}_0^i & : \quad \mu_i = 0 \quad \text{(no source)} \\ \mathcal{H}_1^i & : \quad \mu_i > 0 \quad \text{(contribution of source)}, \end{cases}$$

where the number of tests equal to $n = d$ can be huge for massive data.

3.2. P-values and Benjamini-Hochberg procedure

In the case of n independent test statistics, Benjamini and Hochberg [3] proposed a procedure that controls the FDR at a level $\pi_0 q \leqslant q$, where $\pi_0 = \frac{n_0}{n}$ and n_0 is the number of tests that are true $\mathcal{H}_0$ (when π_0 is unknown, the FDR is controlled at nominal level q):

1. Evaluate the n p-values p_i, $i = 1 \cdots n$.
2. Let $p_{(0)} = 0$ and $p_{(i)}$, $i = 1 \cdots n$, the ordered p-values: $p_{(0)} < p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(n)}$.
3. Define $k = \underset{i}{\text{argmax}} \left(p_{(i)} \leqslant q \frac{i}{n} \right)$
4. Reject $\mathcal{H}_0^{(1)}, \cdots, \mathcal{H}_0^{(k)}$.

To apply the Benjamini-Hochberg (BH) procedure to our case, we consider the p-values $p_i = 1 - F_{\mathcal{H}_0}(X_i) = F_{\mathcal{H}_0}(-X_i)$ where $F_{\mathcal{H}_0}$ is the cumulative density function of the test under the null hypothesis, i.e the cumulative density function of the standard normal distribution. By construction the p-value p_i is uniformly distributed on $[0, 1]$ under the null hypothesis $\mathcal{H}_0^i$ while it is stochastically lower, *i.e.* $\Pr(p_i < t) > t$ for $0 \leqslant t \leqslant 1$, under the alternative hypothesis $\mathcal{H}_1^i$.

3.3. Threshold a Gaussian vector with BH procedure

Benjamini and Yekutieli [1] proposed an extension of the BH procedure to dependent tests under the positive regression dependency on a subset (PRDS) assumption (see [1, p1168] for details). They showed if $\mathbf{\Sigma} \geqslant 0$, *i.e.* if all the matrix entries are non-negative, then $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ is PRDS. We propose to threshold the matched filter output, which is, from (3), a Gaussian vector, with BH procedure under PRDS assumption. This yields the following proposition.

Proposition. *If the matched filter output is Gaussian, $X = H^TY \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, with a non-negative covariance $\mathbf{\Sigma} \geqslant 0$, then the corresponding p-values are PRDS.*

From [1], if a vector is PRDS on any subset with and if f is a monotone function then the vector $Y = (f(X_1), \ldots, f(X_n))$ is also PRDS. This is true in particular for the p-values that are decreasing function of the Gaussian vector X: $p_i = 1 - F(X_i)$, where F is the cumulative density function of the test under the null hypothesis. Thus we state the following proposition.

Proposition. *Applying the BH procedure allows to threshold the matched filter output while controlling the FDR.*

In our case, the matched filter result $X = H^TY$ is PRDS because of (3) and the BH procedure allows to control the FDR at level $q\frac{n_0}{n} \leqslant q$ by thresholding the p-values. Rejecting $\mathcal{H}_0^{(1)}, \cdots, \mathcal{H}_0^{(k)}$ by the BH procedure is equivalent to decide $X_{(1)}, X_{(2)}, \cdots, X_{(k)}$ are significative, *i.e.* $\mu_{(1)} > 0, \mu_{(2)} > 0, \cdots, \mu_{(k)} > 0$, where the notation $X_{(1)}, X_{(2)}, \cdots, X_{(k)}$ stands for the tests corresponding to the ordered p-values $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(k)}$.

3.4. Example

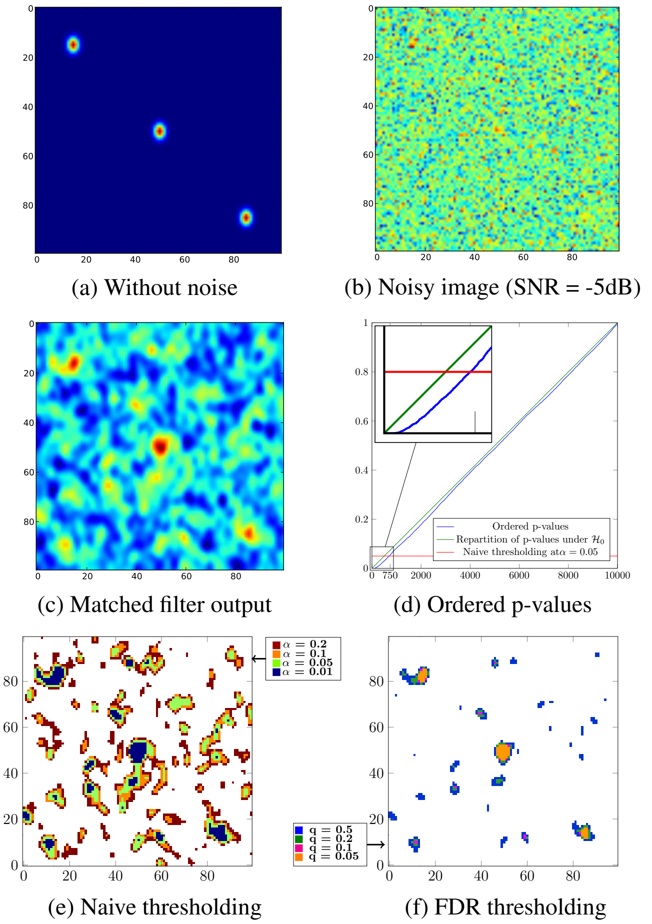
3.4.1. Matched filtering

Below is a simulation of a source detection problem using the FDR control on the matched filter output. Consider a two-dimensional noisy image containing three identical sources of known profile with a -5dB SNR. Figure 1(a) shows three sources without noise in a 100×100 pixels image. Figure 1(b) illustrates the difficulty to detect these sources in presence of Gaussian noise. Figure 1(c) is the matched filter output and figure 1(d) shows the ordered p-values graph computed on the filtered image and the theoretical repartition under $\mathcal{H}_0$. The closer to zero the p-value is, the more likely to belong to a source the corresponding pixel is. These p-values must be thresholded while controlling an error control criterion.

Figures 1(e) and 1(f) compare the performance of naive thresholding at different significance levels α and the performance of the matched filter output thresholding with the BH procedure for different FDR levels.Threshold naively the corresponding p-values at a classical level $\alpha = 0.05$ (model by the red thresholding on figure 1(d)) leads to a large number of false detections (see figure 1(e)). In this case, more than 50% of the selected pixels are actually pure noise pixels. Controlling the FDR seems to be more appropriated to screen the pixels that may be affected by a source while limiting the errors. Moreover the few false detected pixels should be eliminated by source identification dedicated algorithms.

3.4.2. Performances

Among the $n = 10^4$ tests, the proportion of true $\mathcal{H}_0$ is $\pi_0 = 0.97$. We apply BH procedure for different FDR and different noise levels (-5dB $\leqslant$ SNR $\leqslant$ 30dB). For each case, 1000 simulations are performed to measure the rate of false discoveries actually obtained by thresholding the matched filter



(a) Without noise

(b) Noisy image (SNR = -5dB)

(c) Matched filter output

(d) Ordered p-values

(e) Naive thresholding

(f) FDR thresholding

Fig. 1: Example of sparse source detection problem: (a) three source responses (b) noisy observed image (c) matched filter output (d) p-values vs their rank orders (blue curve), theoretical p-value quantiles under the null vs rank order i ($y = i/n$ green curve), $\alpha = 0.05$ threshold for a single test (horizontal red line in the zoom) (e) thresholding using respectively different significative levels α based on a single test, and (f) using the BH procedure for different FDR levels q

output. The BH procedure controls the FDR at $\pi_0 q \leqslant q$. Figure 2 shows the false discoveries rate for a given control q, and the power of this procedure. The actual FDR are close to the control level $\pi_0 q$, and for $q \leqslant 0.2$ the powers are very similar. Therefore the same power can be maintained here with a low false discoveries rate.

4. APPLICATION TO THE GALAXY DETECTION

Finding faint and unresolved (or poorly resolved) objects in massive data (large images, hyperspectral cubes, etc) while controlling FDR is a current issue in astronomy. This problem can actually be recasted in the framework described in the previous section.

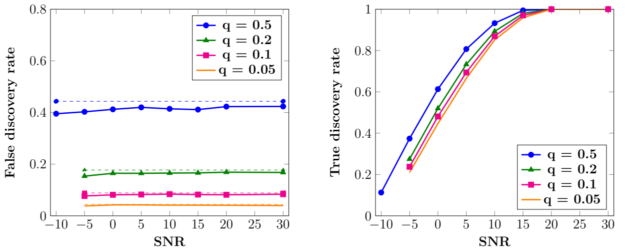


Fig. 2: False (left) and true (right) discoveries rate estimated on 1000 simulations of the noisy image containing three sources for different FDR and different SNR. True discovery is a true $\mathcal{H}_1$ test correctly identified.

4.1. Data description

The Multi Unit Spectroscopic Explorer (MUSE) is a second generation VLT instrument [11]. MUSE has a field of 1×1 arcmin^2 sampled at 0.2×0.2 arcsec^2 and spectra are sampled at 1.25 Å for wavelength ranging from 4800Å to 9300Å. The MUSE instrument is able to record more than 90000 spectra by image. Classical dimensions of a MUSE data cube are $300 \times 300 \times 3600$ pixels. Different kinds of galaxies are expected to be found in the MUSE data cube. Some are very bright or spatially extended ones and easy to detect. The real challenge consists on detecting distant galaxies with a very low intensity and low spatial extension,whose spectral characteristics reduces to a single emission line of a few Å wide. Then the response of such a source should be close to the 3D point spread function (PSF) of the MUSE instrument.

4.2. 3D matched filter

The MUSE PSF spectrally varies, so does the point source response. The matched filter has to include this variability for improving the detection. The intensity profile h_λ is a three-dimensional positive kernel, defined for all of the 3600 spectral bands. The correlation induced by the matched filter involves $13 \times 13 \times 7$ pixels, which is the size of the PSF kernel h_λ.

4.3. 3D thresholding procedure

The matched filter and BH procedure are applied on the MUSE view of the Hubble Deep Field South (HDFS)[1] [12]. In order to improve the tractability of dedicated detection algorithm [13], a first step is proposed, whose objective is to provide a screening of the candidate pixels. Figure 3 actually represents a projection (integrating all wavelengths) of the 3D thresholding procedure result for a FDR set to $q = 0.1$. A lot of regions of low spatial and spectral extension are rejected by the BH thresholding procedure, actually 3.7% of the three-dimensional data cube have been screened. These pixels can now be proposed to the source finder algorithm [13] as candidates for faint galaxy centers. Note that this data cube

[1] All the material is available at http://muse-vlt.eu/science/

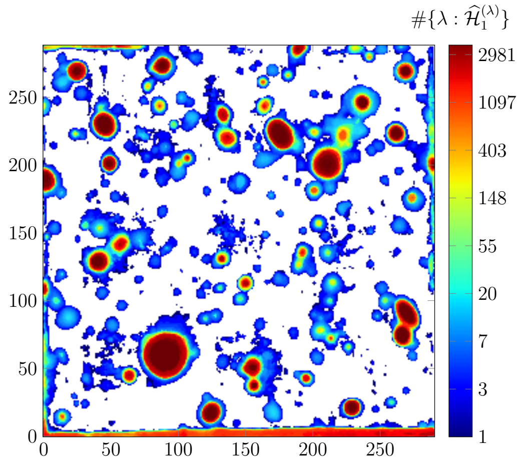


Fig. 3: Projection of the threshold data cube. The color scale indicates the number of activated wavelengths for each spatial position r, *i.e.* the number of spectral components classified in $\mathcal{H}_1$ for each position r (FDR = 0.1).

contains also spatially and spectrally extended bright galaxies which are highlighted by the matched filter, that explains the large dark sources on figure 3. Among these bright galaxies, pixels belonging to sources of low spatial and spectral extension are also detected. The procedure is robust to the spectral variability of the sources.

4.4. Discussion

In the MUSE data case, the proportion π_0 is not known, so the theoretical upper bound $\pi_0 q$ is unknown. Then the BH procedure guarantees a FDR control upper bounded by the chosen value q. As π_0 is expected to be close to one for this application, these bounds are almost the same. Otherwise the proportion π_0 can also be estimated to obtain a sharper bound. Some approaches have been derived in the literature (see for instance [14, 15] and [16]) that guarantee FDR control $\widehat{\pi}_0 q$ in the case of independent tests. Note that in [15], the control is asymptotically guaranteed under local dependency between tests. This is for instance the case of the matched filter output when the source responses are sparse.

Another point is worth being mentioned. For the MUSE data, an independent chi-squared distributed estimator S^2 of the noise variance is provided with the data. Normalizing the MUSE data cube with S leads to a studentized vector Y/S whereas in the previous sections we consider an equivariate normal distributed vector. From [1], Y/S is not PRDS but the BH procedure can be used for controlling FDR provided that $q < 0.5$ which is the case in practice. This result is used for producing the map shown on Figure 3.

5. CONCLUSION

Based on positivity assumption for the matched filter response, we show that applying the BH procedure to threshold the matched filter output allows to control the FDR despite the dependency. We use this result for building a 3D map of pixels that may belong to faint galaxies of low spatial extension in massive hyperspectral data. This map of candidate pixels may actually be used as an input of galaxy detection algorithm [13], allowing to control both the FDR and computational tractability.

REFERENCES

[1] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[2] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[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] C. J Miller, C. Genovese, R. C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, J. Schneider, and A. Moore, "Controlling the false-discovery rate in astrophysical data analysis," *The Astronomical Journal*, vol. 122, no. 6, pp. 3492, 2001.

[5] A. M. Hopkins, C. J. Miller, A.J. Connolly, C. Genovese, R. C. Nichol, and L. Wasserman, "A new source detection algorithm using the false-discovery rate," *The Astronomical Journal*, vol. 123, no. 2, pp. 1086–1094, 2002.

[6] 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.

[7] A. Reiner, D. Yekutieli, and Y. Benjamini, "Identifying differentially expressed genes using false discovery rate controlling procedures," *Bioinformatics*, vol. 19, no. 3, pp. 368–375, 2003.

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

[9] M. Al-Rawi, M. Qutaishat, and M. Arrar, "An improved matched filter for blood vessel detection of digital retinal images," *Computers in Biology and Medicine*, vol. 37, no. 2, pp. 262–267, 2007.

[10] R. Vio, P. Andreani, and W. Wamsteker, "Some good reasons to use matched filters for the detection of point sources in cmb maps," *Astronomy and Astrophysics*, vol. 414, no. 1, pp. 17–21, 2004.

[11] R. Bacon et al., "Probing unexplored territories with MUSE: a second generation instrument for the VLT," in *SPIE 6265*, 2006.

[12] R Bacon, J Brinchmann, J Richard, et al., "The MUSE 3D view of the Hubble Deep Field South," *arXiv preprint arXiv:1411.7667*, 2014.

[13] C. Meillier, F. Chatelain, O. Michel, and H. Ayasso, "Non-parametric bayesian extraction of object configurations in massive data," *IEEE Trans. on Signal Processing*, accepted in February, 2015.

[14] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002.

[15] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 1, pp. 187–205, 2004.

[16] G. Blanchard and É. Roquain, "Adaptive false discovery rate control under independence and dependence," *The Journal of Machine Learning Research*, vol. 10, pp. 2837–2871, 2009.