10.2.3 WP 1-6X2: New point source detection methods

PROJECT: **Studi di Cosmologia**		**WP REF.: 1-6X2**
WP TITLE: **New point source detection methods** **SUB-CONTRACTOR**: **Dipartimento di Matematica, Università Tor Vergata** **START EVENT**: KO **END EVENT**: RF **WP MANAGER**: **Domenico Marinucci**		**Sheet**: 1 **of** 1 **Issue Ref**: 1 **Issue Date**:01/09/2016

1. INPUTS

- *Data* This item can developed primarily out of Planck data; applications to future ground-based or satellite experiments can also eb envisaged.
- *Technical Tools* As detailed below, this part of the project is going to build upon some very recent developments in the mathematical statistics literature, concerning multiple testing on spherical data.
- *Software* Our routines require specific packages for data analysis in the needlet domain. These packages are now well-developed within the *HealPix* paradigm.

2. TASKS

Our main goal for this task is to develop and adjust for the requirements of CMB data analysis the state-of-the-art techniques in a statistical framework for the principled implementation of multiple testing algorithms. As well known, indeed, point-source detection in either the temperature or the polarization domains can be viewed as a typical multiple testing issue, where the possible presence of several thousands sources at different (unknown) locations is simultaneously investigated. There is currently a vast statistical literature on error control and power maximization in these circumstances; this literature, however, has typically been developed for the case where a fixed (albeit huge) number of tests is implemented, for instance for genome-wide arrays. On the other hand, in the CMB community multiple testing has often been implemented by simply considering the threshold significance of possible point sources on one-by-one basis, thus neglecting statistical issues on global error control. Indeed, a number of algorithms have been proposed for these tasks; these solutions have all been shown to perform well in practice, and indeed they have been widely applied to the analysis of Planck data: however, they have all avoided to face the specific challenges of multiple testing, and in particular none of them has been shown to control in any proper statistical way any aggregate statistics such as the classical Family-Wise Error Rate (FWER), False Discovery Proportion (FDP) or False Discovery Rate (FDR).

More specifically, our aim is to implement new techniques to control the False Discovery Rate (FDR) when detecting multiple point sources. We recall that FDR is defined as the proportion of selected peaks which are not point sources; for massive CMB data sets, it seems very reasonable to search procedures aimed at controlling the proportion of pixels wrongly identified as sources, rather than naively pretending that no such misclassification error can occur. The idea we are going to pursue is to exploit recent developments on the probabilistic properties of multi-scale data analysis techniques. In particular, the goal is to develop CMB maps into needlet components; the distribution of peaks for at high frequencies for these components can be derived analytically, and the control of FDR can then be achieved by a statistical procedure called STEM (smoothing and testing of local maxima). Some properties of these procedures for the analysis of spherical data have been investigated very recently in the mathematical statistics literature, but their calibration and implementation on CMB data is still an open issue for research; the approach is also suitable for application on spherical data outside the CMB framework.

One important extra-advantage of the procedure we are going to develop is that it is not simply going to provide a list of candidate point sources (to be cross-cheched with existing catalogues); indeed, for each of them it will also provide an evaluation of the *p*-value or significance as a point source. This means, for instance, that when planning follow-up observational strategy astrophysicists will be able to decide on their own the level of significance for the sources to be included in their own catalogue. This result can be obtained by implementing on *Planck* CMB data some recent theoretical advances on the distribution of critical points and maxima for filtered spherical Gaussian maps.

These tasks have a number of connections and interactions with other topics of the project. Among these, we mention in particular *Foreground Cleaning* and *Delensing*, both located at Sissa Trieste. The importance of point source detection for foreground cleaning need not be explained; for delensing, we plan to further enhance the exploitation of cross-correlation techniques for the reconstruction of the lensing potantial, an issue on which the collaboration with Sissa has been going on quite rapidly in the last few years.

3. OUTPUTS

- A software to implement the Multiple Testing Algorithm for full-sky temperature and polarization data
- A software to implement the algorithm on small patches of the sky observed at very high resolution.

4. TIMELINES

- First Year, first semester: The Multiple Testing Algorithm is implemented on Planck temperature data
- First year, second semester: The Multiple Testing Algorithm is optimized for Planck temperature data
- Second year, first semester: The Multiple Testing Algorithm is implemented on Planck temperature data
- Second year, second semester: The Multiple Testing Algorithm is optimized for Planck Polarization data
- Third Year, first semester: The algorithm is implemented on small sky patches, observed at very high resolution
- Third year, second semester: The algorithm is optimized for small sky patches, observed at very high resolution