

# A Bayesian technique for the detection of point sources in cosmic microwave background maps

F. Argüeso,<sup>1</sup>\* E. Salerno,<sup>2</sup> D. Herranz,<sup>3</sup> J. L. Sanz,<sup>3</sup> E. E. Kuruoğlu<sup>2</sup> and K. Kayabol<sup>2</sup>

<sup>1</sup>*Departamento de Matemáticas, Universidad de Oviedo, 33007 Oviedo, Spain*

<sup>2</sup>*National Research Council of Italy-CNR, ISTI, via G. Moruzzi 1, I-56124 Pisa, Italy*

<sup>3</sup>*Instituto de Física de Cantabria, CSIC-UC, Av. los Castros s/n, 39005 Santander, Spain*

Accepted 2011 January 20. Received 2011 January 7; in original form 2010 September 22

## ABSTRACT

The detection and flux estimation of point sources in cosmic microwave background (CMB) maps is a very important task in order to clean the maps and also to obtain relevant astrophysical information. In this paper we propose a maximum a posteriori (MAP) approach detection method in a Bayesian scheme which incorporates prior information about the source flux distribution, the locations and the number of sources. We apply this method to CMB simulations with the characteristics of the *Planck* satellite channels at 30, 44, 70 and 100 GHz. With a similar level of spurious sources, our method yields more complete catalogues than the matched filter with a  $5\sigma$  threshold. Besides, the new technique allows us to fix the number of detected sources in a non-arbitrary way.

**Key words:** methods: data analysis – techniques: image processing – cosmic background radiation – radio continuum: galaxies.

## 1 INTRODUCTION

The detection and estimation of the intensity of compact objects embedded in a background plus instrumental noise is a problem of interest in many different areas of science and engineering. A classic example is the detection of point-like extragalactic objects – i.e. galaxies – in sub-millimetric astronomy. Regarding this particular field of interest, different techniques have proven useful in the literature. Some of the existing techniques are: the standard matched filter (MF; Nailong 1992), the matched multi-filter (Herranz et al. 2002; Lanz et al. 2010) or the recently developed matched matrix filters (Herranz & Sanz 2008). Other methods include continuous wavelets like the standard Mexican Hat (Sanz et al. 2006) and other members of its family (González-Nuevo et al. 2006). All these filters have been applied to real data of the Cosmic Microwave Background (CMB), like those obtained by the *Wilkinson Microwave Anisotropy Probe* (WMAP) satellite (López-Caniiego et al. 2007) and CMB simulated data (Leach et al. 2008) for the experiment on board the *Planck* satellite (Tauber 2005). Besides, Bayesian methods (BMs) have also been recently developed (Hobson & McLachlan 2003; Carvalho, Rocha & Hobson 2009). A more detailed review on point source detection techniques in microwave and sub-mm astronomy, with a more complete list of references, can be found in Herranz & Vielva (2010).

When an MF or a wavelet is applied to a CMB map in the blind detection case, i.e. when it is assumed that the number of point sources, their positions and fluxes are unknown, the most

common method for detection is based on the well-known idea of thresholding: the maxima of the filtered map above a given threshold are selected and considered as the positions of the sources, so that the number of detected sources is the number of maxima above that threshold. The fluxes are estimated then by using the corresponding estimation formulae with the MF or the wavelet. The value of this threshold remains arbitrary, though a  $5\sigma$  cut is often applied, since it guarantees that under reasonable conditions a few detected sources are spurious. Apart from the arbitrariness of this procedure, the prior knowledge regarding the average number of sources in the surveyed patch, the flux distribution of these sources or other properties are not used, so that useful information is being neglected.

Bayesian detection techniques provide a natural way to take into account all the available information about the statistical distribution of both the sources and the noise. Unfortunately, up to this date only a few works have addressed the problem of detecting extragalactic point sources in CMB data (Hobson & McLachlan 2003; Carvalho et al. 2009). The reason for this is twofold: on the one hand, the statistical properties of extragalactic sources at sub-mm frequencies are still very poorly known. On the other hand, mapping the full posterior probability density of the sources is often very difficult and computationally expensive. These two problems explain, at least partially, the predominance of frequentist over BMs in the literature. Let us consider the previous two problems separately.

The microwave and sub-mm region have been until very recently one of the last uncharted areas in astronomy. Concerning extragalactic sources, this region of the electromagnetic spectrum is where the total number of counts passes from being dominated by radio-loud galaxies to being dominated by dusty galaxies. Although a minimum of the emission coming from extragalactic sources is expected

\*E-mail: argueso@uniovi.es

to occur around 100–300 GHz, they are still considered as the main contaminant of the CMB at small angular scales at these frequencies (Toffolatti et al. 1998; de Zotti et al. 2005). The uncertainties about the number counts at intermediate and low flux, redshift distribution, evolution and clustering properties of this mixed population of objects are large. In most cases this has motivated the use of non-informative priors, which avoid to make adventurous assumptions about the sources but on the other hand miss part of the power of the priors that are based on observations and physical intuition.

But, in spite of what has been said above, our knowledge about the statistical properties of point sources is growing day by day thanks to the new generation of surveys and experiments. In the high-frequency radio regime, *WMAP* observations are in agreement with the de Zotti model (de Zotti et al. 2005; González-Nuevo et al. 2008). Priors for the number density and flux distributions in the range of frequencies  $>5$  GHz are more and more reliable thanks to the information provided by recent surveys such as Combined Radio All-sky Targeted Eight-GHz Survey (CRATES; Healey et al. 2007), the Ryle-Telescope 9C at 15.2 GHz (Taylor et al. 2001; Waldram et al. 2003) or the AT20G survey at 20 GHz (Ricci et al. 2004; Massardi et al. 2008; Mahony et al. 2010). For a recent review on radio and millimetre surveys and their astrophysical implications, see de Zotti et al. (2010). The situation is worse in the far-infrared part of the spectrum, where relatively large uncertainties remain in the statistical properties, the evolution and, above all, the clustering properties of dusty galaxies. Most of the existing dusty galaxy surveys have been carried out in the near and medium infrared with *IRAS*, *ISO* and *Spitzer*, but the waveband from 60 to 500  $\mu\text{m}$  is still virtually *terra incognita*. The only survey of a large area of the extragalactic sky at a wavelength above 200  $\mu\text{m}$  is the one recently carried out by the Herschel pathfinder experiment, the Balloon Large Area Survey Telescope (BLAST; Devlin et al. 2009). In the next few months, however, the luminosity function and the dust-mass function of dusty galaxies in the nearby Universe will be much better understood thanks to the Herschel-ATLAS Survey (Eales et al. 2010), which covers the wavelength range between 110 and 500  $\mu\text{m}$  and has already produced interesting results during the Herschel Science Demonstration Phase (Clements et al. 2010). Thanks to these and the previously mentioned observations, the sub-mm gap is narrowing and our knowledge of galaxy populations in this wave band, albeit far from perfect, is quickly improving.

Apart from the uncertainties on the priors, the other complication that has traditionally deterred microwave astronomers from attempting Bayesian point source detection is computational and algorithmic complexity. Depending on the choice of priors and the likelihood function, the full posterior distribution of the parameters of the sources may be very complex and in most cases it is impossible to obtain maximum a posteriori (MAP) values of the parameters and their associated errors via analytical equations. Numerical sampling techniques such as Monte Carlo Markov Chain (MCMC) methods are required in order to solve the inference problem, but these methods are computationally intensive. It is thus necessary to apply computing techniques specifically tailored for accelerating the convergence and improving the efficiency of the sampling (Feroz & Hobson 2008) and/or to find smart approximations of the posterior near its local maxima (Carvalho et al. 2009). But these enhancements have the cost of increasing dramatically the algorithmic complexity of the detection software, introducing new layers of intricacy in the form not only of additional assumptions and routines, but also of regularization ‘constants’, hidden variables, hyperparameters and selection thresholds that in many cases must be fine-tuned manually in order to be adapted to the specific

circumstances of a given data set. The complexity of the algorithms can rise to almost baroque levels, having a negative effect on the portability of the codes and on the reproducibility of the results.

We propose in this paper a simple strategy based on Bayesian methodology (BM) which incorporates sensible prior information about the source locations, the source fluxes and the source number distribution. With these priors and assuming a Gaussian likelihood, we can obtain an explicit form of the negative log-posterior of the number of sources and their fluxes and positions. Assuming an MAP methodology, we introduce a straightforward top-to-bottom detection algorithm that allows us to determine the number, fluxes and positions of the sources. We give a simple proof that the positions of the sources *must* be located in the local maxima of the matched-filtered image if there is not a significant overlap between sources. The main computational requirement of our algorithm is the solution of a system of non-linear equations. Our method differs from the one presented by Carvalho et al. (2009) in five main points.

- (i) We use a more realistic set of priors for the source number, intensity and location distributions. In particular, our choice of the prior on the locations is also flat but depends on the number of sources  $n$ , which later proves to be decisive for the log-posterior.
- (ii) We obtain an explicit form of the negative log-posterior and an explicit solution of the MAP estimate of the source intensities as the solution of a non-linear system of equations.
- (iii) We prove that, for non-overlapping sources and a Gaussian likelihood, the MAP estimation of the positions of the sources is given by the location of the local maxima of the matched-filtered images.
- (iv) Since we are interested only in point sources, we fix the size parameter of the objects to be detected.
- (v) We can also find the MAP solution for the number of sources present in the images with a simple top-to-bottom search strategy. We do not need to resort to costly evaluations of the Bayesian evidence.

The layout of the paper is as follows. In Section 2 we present the method and derive the corresponding posterior which includes the data likelihood and the priors. In Section 3 we apply the method to CMB simulations with the characteristics of the Planck radio channels (from 30 to 100 GHz, where the number count priors are most reliable) and compare it with the standard procedure of using an MF with a  $5\sigma$  threshold. The main results are also presented in Section 3. The conclusions are given in the final section.

## 2 METHODOLOGY

In a region of the celestial sphere, we suppose to have an unknown number  $n$  of radio sources that can be considered as point-like objects if compared to the angular resolution of our instruments. This means that their actual size is smaller than our smallest resolution cell. The emission of these sources is superimposed to a radiation  $f(x, y)$  coming from diffuse or extended sources. In our particular case this radiation is the CMB plus foreground radiation. A model for the emission as a function of the position  $(x, y)$  is

$$\tilde{d}(x, y) = f(x, y) + \sum_{\alpha=1}^n a_{\alpha} \delta(x - x_{\alpha}, y - y_{\alpha}), \quad (1)$$

where  $\delta(x, y)$  is the 2D Dirac delta function, the pairs  $(x_{\alpha}, y_{\alpha})$  are the locations of the point sources in our region of the celestial sphere and  $a_{\alpha}$  are their fluxes. We observe this radiation through an instrument, with beam pattern  $b(x, y)$ , and a sensor that adds a

random noise  $n(x, y)$  to the signal measured. Again, as a function of the position, the output of our instrument is

$$d(x, y) = \sum_{\alpha=1}^n a_{\alpha} b(x - x_{\alpha}, y - y_{\alpha}) + (f * b)(x, y) + n(x, y), \quad (2)$$

where the point sources and the diffuse radiation have been convolved with the beam. In our application, we are interested in extracting the locations and the fluxes of the point sources. We thus assume that the fluxes of the point sources are sufficiently above the level of the rest of the signal plus the noise, and consider the latter as just a disturbance superimposed to the useful signal. If  $\epsilon(x, y)$  is the sum of the diffuse signal plus the noise, model (2) becomes

$$d(x, y) = \sum_{\alpha=1}^n a_{\alpha} b(x - x_{\alpha}, y - y_{\alpha}) + \epsilon(x, y). \quad (3)$$

If our data set is a discrete map of  $N$  pixels, the above equation can easily be rewritten in vector form, by letting  $\mathbf{d}$  be the lexicographically ordered version of the discrete map  $d(x, y)$ ,  $\mathbf{a}$  be the  $n$ -vector containing the positive source fluxes  $a_{\alpha}$ ,  $\boldsymbol{\epsilon}$  be the lexicographically ordered version of the discrete map,  $\epsilon(x, y)$ , and  $\boldsymbol{\phi}$  be an  $N \times n$  matrix whose columns are the lexicographically ordered versions of  $n$  replicas of the map  $b(x, y)$ , each shifted on one of the source locations. Equation (3) thus becomes

$$\mathbf{d} = \boldsymbol{\phi} \mathbf{a} + \boldsymbol{\epsilon}. \quad (4)$$

Looking at equations (3) and (4), we see that, if the goal is to find locations and fluxes of the point sources, our unknowns are the number  $n$ , the list of locations  $(x_{\alpha}, y_{\alpha})$ , with  $\alpha = 1, \dots, n$  and the vector  $\mathbf{a}$ . It is apparent that, once  $n$  and  $(x_{\alpha}, y_{\alpha})$  are known, matrix  $\boldsymbol{\phi}$  is perfectly determined. Let us then denote the list of source locations by the  $n \times 2$  matrix  $\mathbf{R}$ , containing all their coordinates. If we want to adopt a Bayesian strategy to solve our problem, we must be able to write the posterior probability density of our unknowns. A suitable estimation criterion must then be chosen.

## 2.1 Posterior

By the Bayes rule, the posterior we are looking for has the following form:

$$p(n, \mathbf{R}, \mathbf{a} | \mathbf{d}) \propto p(\mathbf{d} | n, \mathbf{R}, \mathbf{a}) p(n, \mathbf{R}, \mathbf{a}) \quad (5)$$

where  $p(\mathbf{d} | n, \mathbf{R}, \mathbf{a})$  is the likelihood function, derived from our data model (4). To find the prior density  $p(n, \mathbf{R}, \mathbf{a})$  we need to make a number of assumptions. Let us first observe that, in principle, both  $\mathbf{R}$  and  $\mathbf{a}$  depend on  $n$ , through the number of their elements. On the other hand, we can safely assume that, once  $n$  is fixed, the fluxes  $\mathbf{a}$  of the sources are independent of their locations. These assumptions lead us to write

$$p(n, \mathbf{R}, \mathbf{a}) = p(\mathbf{R}, \mathbf{a} | n) p(n) = p(\mathbf{R} | n) p(\mathbf{a} | n) p(n). \quad (6)$$

This expression is valid when we consider extragalactic point sources, whose fluxes are not related to their positions. This will be the case in this paper.

## 2.2 Likelihood function

As mentioned above, the likelihood function derives from the physics associated with the assumed data model. In general, unfortunately, a data model of type (2) is difficult to describe statistically. We are going to assume from now on that  $\boldsymbol{\epsilon}$  is a random Gaussian field with zero mean and known covariance matrix  $\boldsymbol{\xi}$ . This is true

if we only consider the CMB and the instrumental noise, excluding other foregrounds. In this paper we will deal with zones of the sky where the foreground contribution is not important or where the foregrounds have been conveniently removed by component separation techniques. The likelihood is thus

$$p(\mathbf{d} | n, \mathbf{R}, \mathbf{a}) \propto \exp \left[ -\frac{(\mathbf{d} - \boldsymbol{\phi} \mathbf{a})^t \boldsymbol{\xi}^{-1} (\mathbf{d} - \boldsymbol{\phi} \mathbf{a})}{2} \right]. \quad (7)$$

Observe that the negative of the exponent in (7) is in any case the squared  $\boldsymbol{\xi}^{-1}$ -norm fit of the reconstructed data to the measurements, and this always carries information about the goodness of our estimate. However, if the Gaussian assumption is not verified, function (7) is not the likelihood of our parameters, and when it is introduced in (5), we do not obtain the posterior distribution we are looking for. In our simulations we will also include the confusion noise due to faint extragalactic sources. This confusion noise is not Gaussian but as we will see later, it does not hamper the detections, since its standard deviation is much lower than that of the CMB plus the instrumental noise for the frequencies considered in this paper. For the sake of simplicity, we will defer to further papers the treatment of the more general case which includes the foregrounds.

## 2.3 Prior on source locations

A priori, it is reasonable to assume that all the different combinations of  $n$  distinct locations occur with the same probability. Then function  $p(\mathbf{R} | n)$  in equation (6) can be written as

$$p(\mathbf{R} | n) = \frac{n!(N - n)!}{N!}, \quad (8)$$

since  $N!/(n!(N - n)!)$  is the number of possible distinct lists of  $n$  locations in a discrete  $N$ -pixel map. This assumption is based on the fact that the sources considered in this paper are spatially uncorrelated.

## 2.4 Prior on fluxes

Experimentally, it has been found that the fluxes of the strongest sources are roughly distributed as a negative power law, with exponent  $\gamma$ . Conversely, the weak sources have fluxes that are roughly uniformly distributed. To include these two behaviours into a single formula, one should first discriminate in some way between weak and strong sources. This can be done empirically by establishing a sort of threshold  $a_0$  on the fluxes and a conditional prior with the form of the Generalized Cauchy Distribution (Rider 1957):

$$p(\mathbf{a} | n) \propto \prod_{\alpha=1}^n \left[ 1 + \left( \frac{a_{\alpha}}{a_0} \right)^p \right]^{-\frac{\gamma}{p}}, \quad (9)$$

with  $p$  a positive number. Distribution (9) obviously assumes that the fluxes of the different sources are mutually independent. This prior clearly shows the behaviour required for strong and weak sources. In order to work with non-dimensional quantities, we define  $x_{\alpha} = a_{\alpha}/a_0$ ; we also assume that we will detect point sources above a minimum flux  $a_m$  that leads to the following normalized distribution:

$$p(x | n) = \prod_{\alpha=1}^n \frac{p}{B \left( \frac{1}{1+x_m^p}; \frac{\gamma-1}{p}, \frac{1}{p} \right)} (1 + x_{\alpha}^p)^{-\frac{\gamma}{p}}, \quad x_{\alpha} \in [x_m, \infty) \quad (10)$$

where  $B$  is the incomplete beta function and  $x_m = a_m/a_0$ . In the next section, we determine the values of  $a_0$ ,  $p$  and  $\gamma$  by fitting this formula to the point source distribution given by the de Zotti counts model (de Zotti et al. 2005).

## 2.5 Prior on the number of sources

We need to establish a discrete probability distribution that expresses the probability of a number of occurrences in a fixed domain once their average density is known, their locations in the domain are mutually independent and no pair of sources can occur in the same location. All these assumptions seem reasonable when applied to the configurations of the point sources in the celestial sphere, at least for radio-frequencies (Argüeso, González-Nuevo & Toffolatti 2003; González-Nuevo, Toffolatti & Argüeso 2005). Assuming a continuous map domain, all the requirements mentioned are satisfied by the Poisson distribution. Strictly speaking, we have a discrete  $N$ -pixel map, so a binomial distribution should be used, but if  $N$  is not too small, a Poisson distribution should model correctly the probability of having  $n$  occurrences of point sources. The prior on  $n$  appearing in (6) is thus

$$p(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad (11)$$

where  $\lambda$ , the intensity of the Poisson variable, is the expected number of sources in the map at hand. The value of  $\lambda$  will depend on the flux detection limit  $a_m$ , the size of the map and the wavelength of the observation.

## 2.6 An explicit expression for the negative log-posterior

If we multiply all the factors which appear in (5) and (6) and calculate the negative log-posterior, we find (apart from additive constants)

$$\begin{aligned} L(n, \mathbf{R}, \mathbf{x}) = & \frac{1}{2} (\mathbf{x}' \mathbf{M} \mathbf{x} - 2\mathbf{e}' \mathbf{x}) - \log(N - n)! - n \log(\lambda) \\ & - n \log(p) + n \log B \left( \frac{1}{1 + x_m^p}, \frac{\gamma - 1}{p}, \frac{1}{p} \right) \\ & + \frac{\gamma}{p} \sum_{\alpha=1}^n \log(1 + x_\alpha^p), \end{aligned} \quad (12)$$

with  $\mathbf{M} = a_0^2 \boldsymbol{\Phi}' \boldsymbol{\xi}^{-1} \boldsymbol{\Phi}$  and  $\mathbf{e} = a_0 \boldsymbol{\Phi}' \boldsymbol{\xi}^{-1} \mathbf{d}$ . The correlation matrix  $\boldsymbol{\xi}$  is computed by using the  $C_s$  obtained from the *WMAP* 5-yr maps (Nolta et al. 2009) and adding the instrumental noise. We assume that we know  $a_0, p, x_m, \gamma$  and  $\lambda$ , in fact we calculate them by using the de Zotti counts model (de Zotti et al. 2005). Therefore, the unknowns are: the normalized fluxes  $\mathbf{x}$ , the number of point sources  $n$  and the positions of the point sources through the matrix  $\boldsymbol{\Phi}(\mathbf{R})$ .

Let us now examine the structure of function (12). The first term comes from the likelihood, and obviously decreases as much as our solution fits the data. The second term takes into account the prior for the source configuration and penalizes large values of  $n$ . The third term comes from the prior on the number of sources and, depending on the value of  $\lambda$ , favours ( $\lambda > 1$ ) or disfavors ( $\lambda < 1$ ) the increase of the number of sources. The subsequent terms come from the prior on the source fluxes conditioned to  $n$ , the last one introduces an additional cost as soon as a new source is added to the solution. If  $N \gg \lambda$  and  $N \gg n$ , what is typical in CMB maps, it can be proven by using Stirling's approximation that the second term is the dominant one coming from the priors. In the next section, we will analyse with simulations the contribution of each particular term.

## 2.7 Maximum a posteriori solution

Formula (12) includes all the information about the positions, fluxes and number of sources. In order to obtain concrete results, we will

choose the values of  $\mathbf{R}$ ,  $\mathbf{x}$  and  $n$  which maximize the posterior. This choice will be justified by means of the simulations and results that we will present in the next section.

Therefore, regarding the flux we minimize (12) with respect to  $\mathbf{x}$ , by taking the derivative and equating to zero and we obtain

$$\sum_{\beta=1}^n M_{\alpha\beta} x_\beta - e_\alpha + \frac{\gamma x_\alpha^{p-1}}{1 + x_\alpha^p} = 0. \quad (13)$$

By solving (13) numerically we would obtain the estimator of  $\mathbf{x}$  which yields the maximum posterior probability. However, we know neither the number of sources nor their positions. In order to determine the positions, we assume that the point sources are in the local maxima of  $\mathbf{e}$ , which is the matched-filtered map of the original data. In the following, we will show that this assumption can be safely adopted, since the minima of  $L$  must be in the maxima of the matched-filtered map (we also remark that the MF is not introduced ad hoc, but it appears naturally as a part of the formalism).

For simplicity, let us assume that we have only one source, in this case the terms of  $L$  which depend on the flux can be written as

$$L(x) = \frac{M_{11} x^2}{2} - e_1 x + \frac{\gamma}{p} \log(1 + x^p), \quad (14)$$

where  $M_{11}$  and  $e_1$  are the corresponding values of  $\mathbf{M}$  and  $\mathbf{e}$  at the pixel supposedly occupied by the point source. If we take the derivative of (14) with respect to  $x$  and equate to zero, we obtain the following equation for  $\hat{x}$ , the estimator of  $x$ :

$$M_{11} x + \frac{\gamma x^{p-1}}{1 + x^p} = e_1 \Rightarrow \hat{x} = \hat{x}(e_1). \quad (15)$$

If we substitute the last expression in equation (14), we can write the expression for the negative log-posterior  $L(\hat{x}(e_1))$

$$L(\hat{x}(e_1)) = -\frac{M_{11} \hat{x}^2}{2} - \frac{\gamma \hat{x}^p}{1 + \hat{x}^p} + \frac{\gamma}{p} \log(1 + \hat{x}^p). \quad (16)$$

By taking the derivative of this formula with respect to  $e_1$ , we finally find

$$\frac{dL}{de_1} = \frac{dL}{d\hat{x}} \left( \frac{d\hat{x}}{de_1} \right)^{-1} = -\hat{x}(e_1), \quad (17)$$

where we have calculated the derivative in (15). Since the estimated value of the flux must be positive, this expression shows that the negative log-posterior at the estimated value of  $x$  decreases with  $e_1$ , so it is minimum at the highest value of  $e_1$ , i.e. at the maximum of the matched-filtered map. Therefore, the posterior, calculated at the estimated flux value, is maximum when we assume that the point source is at a peak of the map. This conclusion is valid if we have more than one source, provided that there is no overlap between sources, i.e. the areas where the individual images of all the sources are non-zero must be completely disjoint, because in this case each source can be treated individually.

In order to determine the number of sources, we sort these local peaks from top to bottom and solve (13) successively adding a new source. At the same time, we calculate the negative log-posterior (12) and choose the number  $n$  of sources which produce the minimum value of (12). In this way we have constructed an objective stopping criterion which yields, by combining (12) and (13), the number of sources and their fluxes which maximize the posterior. In the next section, we apply the method explained above to the detection and flux estimation of point sources in CMB maps.

In order to compare this technique with a standard method, we also calculate the local peaks above a certain threshold, for instance a  $5\sigma$  threshold, and solve (13) with  $\gamma = 0$ , that amounts to using an MF, i.e. a maximum likelihood estimator.

**Table 1.** Values of the power-law exponent  $\gamma$  and the flux  $a_0$  in Jy, as obtained by fitting the De Zotti counts.  $\lambda$  is the average number of point sources above 0.25 Jy in the considered patches.

Frequency (GHz)	$\gamma$	$a_0$	$\lambda$
30	2.90	0.19	0.69
44	2.87	0.15	0.53
70	2.87	0.15	0.49
100	2.87	0.15	0.47

### 3 SIMULATIONS AND RESULTS

#### 3.1 The simulations

In order to check the performance of the new technique, we have carried out simulations including CMB, instrumental noise and point sources. The simulations have the characteristics of the 30, 44, 70 and 100 GHz channels of the *Planck* satellite: pixel size, beamwidth and instrumental noise.<sup>1</sup> The simulations are flat patches of  $32 \times 32$  pixels (30 and 44 GHz) and  $64 \times 64$  pixels (70 and 100 GHz), so that the size of each patch is  $3.66 \times 3.66 \text{ deg}^2$ . In order to avoid border effects, we simulate patches of four times this size and keep the central part for our analysis. The small size of the simulations allows us to do our calculations in a fast way. We perform 1000 simulations for each channel.

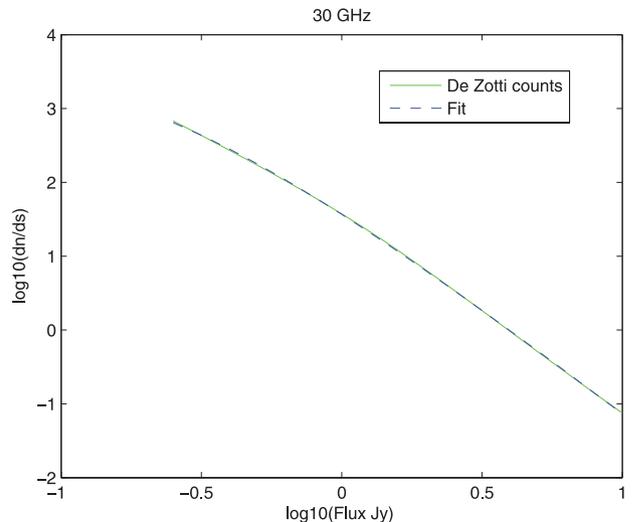
The CMB maps have been generated by using the power spectrum, the  $C_\ell$ s, that produces the best fit to the *WMAP* 5-y maps (Nolta et al. 2009); we have also added the instrumental noise of the 30, 44, 70 and 100 GHz channels of the *Planck* satellite. Finally, we have simulated point sources by taking into account the flux distribution predicted by the de Zotti model (de Zotti et al. 2005). We have included the faint tail of the de Zotti distribution, simulating point sources from 0.01 mJy on. In this way, we have considered the confusion noise due to unresolved point sources. The standard deviation of this confusion noise is much lower than that of the CMB plus instrumental noise in these channels.

For each simulation we consider the negative log-posterior given by (12). In this equation, we see several magnitudes,  $\gamma$ ,  $a_0$ ,  $a_m$  and  $\lambda$ , which depend on the frequency. By fitting the number counts given in the de Zotti model by (9), we calculate  $\gamma$  and  $a_0$ . We have taken  $p = 1$  in (9), since the goodness-of-fit obtained by changing  $p$  is not better than that of the particular case  $p = 1$ . The parameter  $\lambda$  is the average number of sources per patch and  $a_m$  is the minimum flux that we consider in our detection scheme, we have chosen  $a_m = 0.25$  Jy, the typical rms deviation of the CMB plus noise maps at the frequencies we consider. The values of  $\gamma$ ,  $a_0$  and  $\lambda$  are shown in Table 1 for the different frequencies. In Fig. 1 we show as an example the fit to our extended power law (9) in the case of the 30 GHz channel. It is clear that the extended power law fits very well the counts predicted by the de Zotti model. The value of  $\chi^2$  is  $(2 - 3) \times 10^{-3}$  giving probabilities very close to 1.

#### 3.2 Discussion on the performance of the algorithm

For each simulation we calculate  $\mathbf{M} = a_0^2 \Phi' \xi^{-1} \Phi$  and  $\mathbf{e} = a_0 \Phi' \xi^{-1} \mathbf{d}$ . We obtain  $\xi^{-1}$  from the *WMAP*  $C_\ell$ s, taking into account the effects of the pixel and the beam windows and the correspond-

<sup>1</sup> For details on the Planck instrumental and scientific performance, see the Planck web site <http://www.rssd.esa.int/index.php?project=PLANCK>.

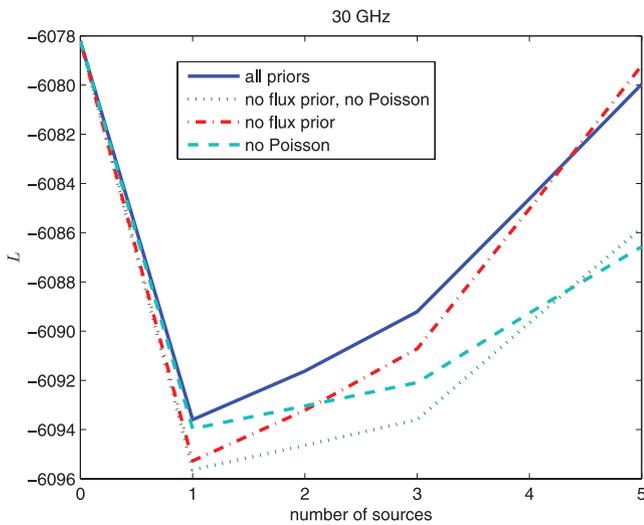


**Figure 1.**  $\log_{10}$  of the differential counts plotted against the flux for the de Zotti model (solid line) and the fit to the extended power law given by (9) with the parameters of Table 1 (dashed line). The two lines are nearly indistinguishable.

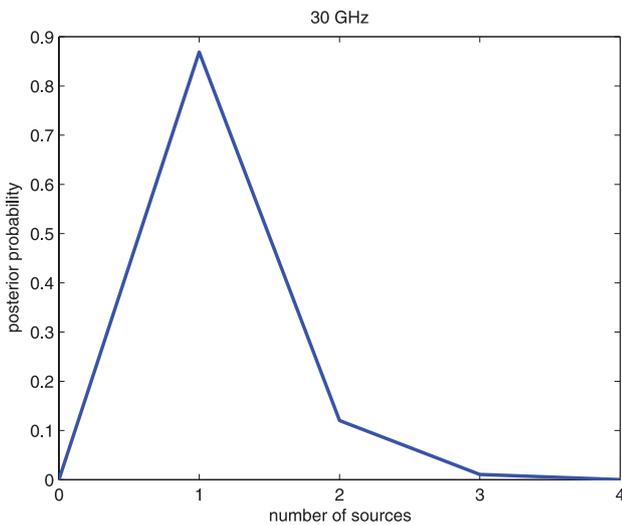
ing noise levels for each channel. We calculate the estimated fluxes  $\hat{x}_\alpha$  by solving (13) as explained in the previous section: we select the maxima of  $\mathbf{e}$  above a certain threshold (we choose a  $1\sigma$  threshold so that we have a suitable number of peaks) and we perform a top to bottom strategy, i.e. we sort the local maxima downwards from higher to lower values and starting from the highest peak we solve (13) including in each new iteration a new local maximum. At the same time, we calculate (12) and stop the iterations when we find the minimum value of the negative log-posterior. In this way, we obtain the source fluxes and the number of sources that maximize the log-posterior. We also calculate the local peaks of  $\mathbf{e}$  above a  $5\sigma$  threshold, a standard detection method, and calculate the source flux by solving (13) with  $\gamma = 0$ ; this is equivalent to using an MF with a  $5\sigma$  threshold. Our intention is to compare the BM, with prior information and a natural stopping criterion, and the standard MF.

According to our simulations, the fundamental contributions to the posterior come from the likelihood (7) and the prior on source locations (8). The other terms also contribute, but as can be seen in Fig. 2, where we show the negative log-posterior plotted against the number of sources for a particular simulation, their influence is not so important. The likelihood tends to increase the number of detected sources, over-fitting the data and the prior (8) tends to decrease the number of sources. The combination of (7) and (8) fixes the most probable number of point sources, though the other two terms, although less important, can have some influence. This shows the robustness of the method with respect to small changes in the parameters of priors (10) and (11).

We also raise the question whether the estimated number of point sources gives us a clearly higher posterior probability, i.e.  $\propto \exp(-\log L)$  than other close numbers. The answer can be seen in Fig. 3, where we show, as an example, the normalized posterior probability plotted against the number  $n$  of point sources for a given simulation with one real source. The probability is clearly peaked at the estimated number of sources, which is the real number of point sources in this case. In our simulations we observe that the posterior probability is always strongly peaked around the estimated number of sources.



**Figure 2.** Negative log-posterior against the number of detected sources for a simulation at 30 GHz. We have included in the posterior: all the priors (blue solid line), all the priors but the source flux distribution (dash-dotted line), all the priors but the Poisson source number distribution (dashed line), and finally we have excluded both the *Poisson* and the source flux distribution (dotted line).



**Figure 3.** Posterior probability against the number of detected point sources for a simulation at the 30 GHz channel with one real source.

We analyse 1000 simulations for each of the considered Planck channels: we calculate the contamination (the number of detected spurious sources over the total number of detected sources above a given flux), the completeness (the number of real detected sources over the number of simulated sources above a given flux) and the average of the absolute value of the relative error of the estimated flux with respect to the real flux (reconstruction error). We count a detected source as real when there is a real simulated source at a distance no longer than two pixels from the detected one; this distance is the position error. This real source must have a flux equal or higher than 0.20 Jy, to fix a threshold close to the  $1\sigma$  level of the CMB plus noise map. The same conditions are required for the MF.

In order to give the uncertainty in the flux, derived from our Bayesian approach, we will obtain a 95 per cent confidence interval

associated with our probability distribution

$$P(x) \propto \exp(-L(x)), \quad (18)$$

where  $L(x)$  is given by equation (14). This will be called the estimation error. We can also calculate the expectation value of the flux from this distribution; this value can be compared with the most probable value (i.e. our estimated flux).

### 3.3 Results

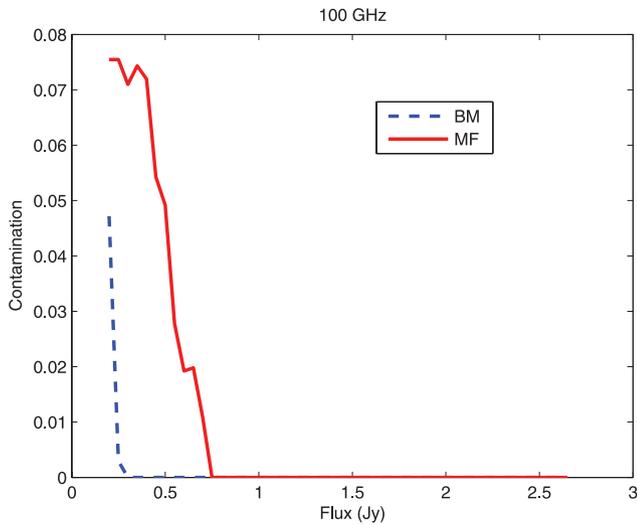
Taking into account the above considerations, we have applied our algorithm to the simulations described in Section 3.1 and obtained the following results.

At the 30 GHz channel we have 4.9 per cent contamination above 0.2 Jy with the BM, and 0.7 per cent with the MF. However, the completeness is much better for the BM, 64 per cent, than for the MF, 22 per cent. From 0.7 Jy on, we do not have any spurious source (BM) and the completeness is 99 per cent. For the MF there are no spurious sources from 0.25 Jy on, but the completeness at 0.7 Jy is only 75 per cent. With regard to the average value of the absolute value of the relative error (reconstruction error), when we calculate this error in flux intervals of 0.1 Jy, we obtain similar values for the BM and the MF. For instance, we obtain errors below 15 per cent from 0.6 Jy on and below 10 per cent from 1 Jy on for both methods. Only 14 per cent of the sources have a reconstruction error on the position of 1 pixel and only 3 per cent have a higher error. These results are nearly the same for the BM and the MF.

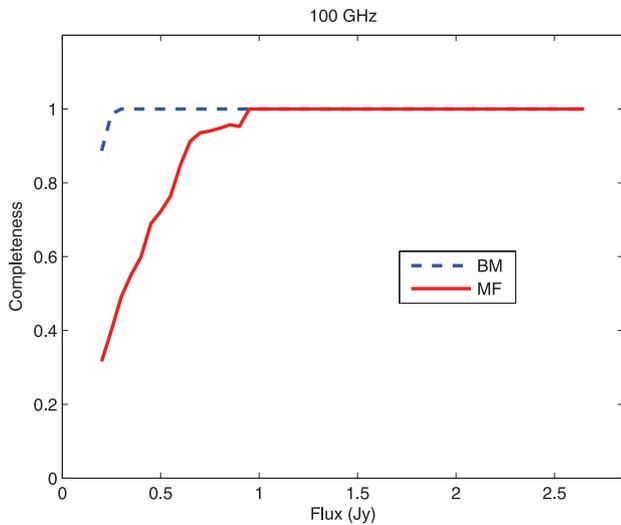
At the 44 GHz channel we have 6.5 per cent contamination for fluxes higher than 0.2 Jy with the BM, and 4 per cent with the MF. The completeness is 37 per cent for the BM and 11 per cent for the MF. From 0.8 Jy on we do not have any spurious source (BM) and the completeness is 100 per cent. For the MF there are no spurious sources from 0.30 Jy on, but the completeness at 0.8 Jy is only 70 per cent. As in the 30 GHz case, we obtain similar average values of the absolute value of the relative error for both methods. For instance, we obtain errors below 15 per cent from 0.90 Jy on in both cases. 15 per cent of the sources have a reconstruction error on the position of 1 pixel and only 3 per cent have a higher error. These results are nearly the same for the BM and the MF.

At the 70 GHz channel, we have 3.2 per cent of spurious sources for fluxes higher than 0.2 Jy with the BM, and 1 per cent with the MF. The completeness is 45 per cent for the BM and 19 per cent for the MF. From 0.45 Jy on we do not have any spurious source (BM) and the completeness is 96 per cent. For the MF there are no spurious sources from 0.30 Jy on, but the completeness at 0.45 Jy is only 46 per cent. As in the cases above, we obtain similar average values of the absolute value of the relative error for both methods. For instance, we obtain errors below 15 per cent from 0.60 Jy on in both cases. 9 per cent of the sources have a reconstruction error on the position of 1 pixel and only 1 per cent have a higher error. These results are similar for the BM and the MF.

At the 100 GHz channel we have 4.7 per cent of spurious sources for fluxes higher than 0.2 Jy with the BM, and 7.5 per cent with the MF. The completeness is 89 per cent for the BM and 32 per cent for the MF. From 0.3 Jy on we do not have any spurious source (BM) and the completeness is 100 per cent. For the MF there are no spurious sources from 0.75 Jy on and the completeness at 0.75 Jy is 94 per cent. As in the above cases, we obtain similar average values of the absolute value of the relative error for both methods. For instance, we obtain errors below 10 per cent from 0.5 Jy in both cases. 1 per cent of the sources have a reconstruction error on



**Figure 4.** Contamination plotted against the flux for the BM and the MF (100 GHz).

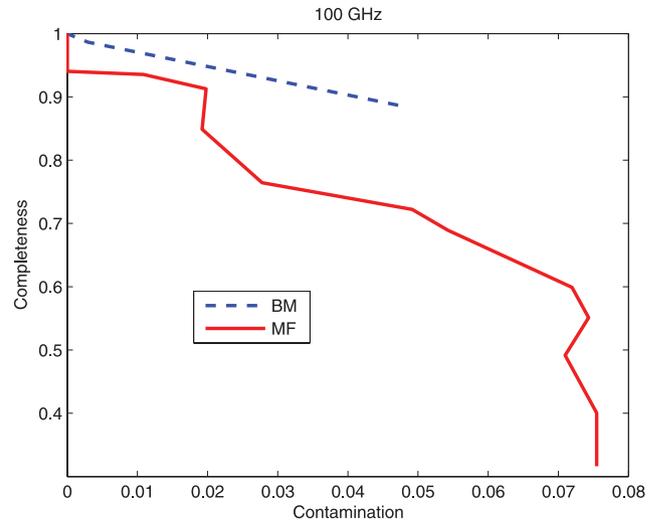


**Figure 5.** Completeness plotted against the flux for the BM and the MF (100 GHz).

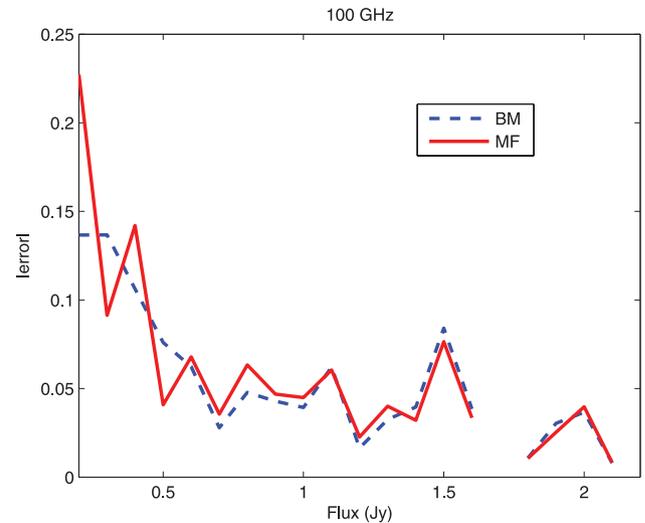
the position of 1 pixel and there are no sources with a higher error. These results are similar for the BM and the MF.

In order to visualize these results, we have plotted for the 100 GHz channel the contamination (integrated contamination) against the flux in Fig. 4, the completeness (integrated completeness) against the flux in Fig. 5, the completeness against the contamination in Fig. 6, the average value of the absolute value of the relative error in Fig. 7 and, finally, the estimated flux against the real flux in Fig. 8. It is clear that for a given value of the contamination the completeness is higher for the BM than for the MF. Although the results are similar at all the studied frequencies, we have chosen the 100 GHz channel in order not to complicate unnecessarily the figures.

In Fig. 9, we plot the expectation value of the flux against the estimated flux for the 100 GHz channel. The expectation value is nearly the same as the most probable value. We also plot the 95 per cent confidence intervals. In this way, we have an idea of the uncertainty of our estimates; this confidence interval is  $\approx 0.20$  Jy (estimation error). The results at other frequencies are similar.



**Figure 6.** Completeness plotted against contamination for the BM and the MF (100 GHz).



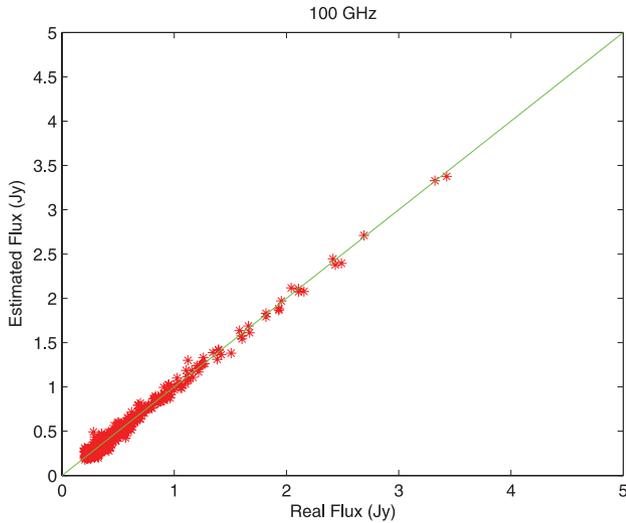
**Figure 7.** Average value of the absolute value of the relative error plotted against the flux for the BM and the MF (100 GHz). We can see in the plot the low values of the error for both methods.

## 4 CONCLUSIONS

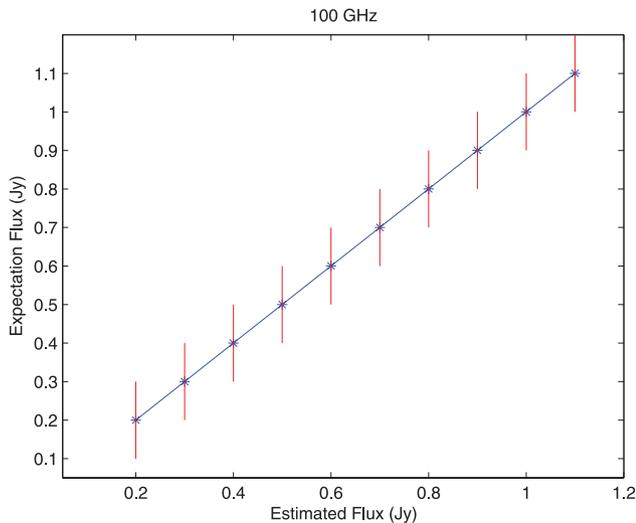
In this paper we propose a new strategy based on BM that can be applied to the blind detection of point sources in CMB maps. The method incorporates three prior distributions: a uniform distribution (8) on the source locations, an extended power law on the source fluxes (10) and a Poisson distribution on the number of point sources per patch (11). Together with a Gaussian likelihood, these priors produce the negative log-posterior (12).

We minimize this negative log-posterior with respect to the source fluxes in order to estimate them. At the same time, we show that the detected sources must be in the peaks of the matched-filtered maps. Finally, we choose the number of point sources which minimizes (12) for the estimated fluxes.

In this way, we give a non-arbitrary method to select the number of point sources. Finally, to check the performance of this technique, we carry out flat CMB simulations for the Planck channels from 30 to 100 GHz. For simplicity, we have excluded the foregrounds in our simulations, assuming that we are considering zones of the sky



**Figure 8.** Estimated flux against real flux for the BM (100 GHz). We have plotted the straight line  $y = x$  for comparison.



**Figure 9.** Expectation value of the flux against estimated flux. The 95 per cent confidence intervals are also plotted (100 GHz).

which have been cleaned by the application of component separation methods. However, we have included the confusion noise due to unresolved point sources in our simulations.

We compare our Bayesian strategy with the application of an MF with a standard  $5\sigma$  threshold. We calculate the completeness, the completeness and the relative error for both methods. Though the percentage of spurious sources is a little higher for the BM at low fluxes  $\simeq 0.2\text{--}0.3$  Jy, the completeness is much better, allowing us to obtain catalogues with a 99 per cent completeness and no spurious sources from 0.7 Jy (30 GHz), 0.8 Jy (44 GHz), 0.55 Jy (70 GHz) and 0.3 Jy (100 GHz) on. The reconstruction errors in the estimated fluxes are similarly low for both methods.

## ACKNOWLEDGMENTS

The authors acknowledge partial financial support from the Spanish Ministerio de Ciencia e Innovación project AYA2007-68058-C03-02 and from the joint CNR-CSIC research project 2008-IT-0059.

KK was supported by the Italian Space Agency, ASI, under the programme on Cosmology and Fundamental Physics, and by the TRIL programme at the Abdus Salam International Centre of Theoretical Physics, through a specific collaboration agreement with ISTI-CNR. ES and EEK acknowledge support from ASI through ASI/INAF Agreement I/072/09/0 for the Planck LFI Activity of Phase E2. We also thank J. González-Nuevo for useful comments and help. We also wish to thank the referee for his comments that have helped us to significantly improve the quality of this work.

## REFERENCES

- Argüeso F., González-Nuevo J., Toffolatti L., 2003, *ApJ*, 598, 86  
 Carvalho P., Rocha G., Hobson M. P., 2009, *MNRAS*, 393, 681  
 Clements D. L. et al., 2010, *A&A*, 518, L8  
 de Zotti G., Massardi M., Negrello M., Wall J., 2010, *Astron. Astrophys. Rev.*, 18, 1  
 de Zotti G., Ricci R., Mesa D., Silva L., Mazzotta P., Toffolatti L., González-Nuevo J., 2005, *A&A*, 431, 893  
 Devlin M. J. et al., 2009, *Nat*, 458, 737  
 Eales S. et al., 2010, *PASP*, 122, 499  
 Feroz F., Hobson M. P., 2008, *MNRAS*, 384, 449  
 González-Nuevo J., Toffolatti L., Argüeso F., 2005, *ApJ*, 621, 1  
 González-Nuevo J., Argüeso F., López-Caniego M., Toffolatti L., Sanz J. L., Vielva P., Herranz D., 2006, *MNRAS*, 369, 1603  
 González-Nuevo J., Massardi M., Argüeso F., Herranz D., Toffolatti L., Sanz J. L., López-Caniego M., de Zotti G., 2008, *MNRAS*, 384, 711  
 Healey S. E., Romani R. W., Taylor G. B., Sadler E. M., Ricci R., Murphy T., Ulvestad J. S., Winn J. N., 2007, *ApJS*, 171, 61  
 Herranz D., Sanz J. L., 2008, *IEEE J. Selected Topics Signal Process.*, 5, 727  
 Herranz D., Vielva P., 2010, *IEEE Signal Processing Magazine*, 27, 67  
 Herranz D., Sanz J. L., Hobson M. P., Barreiro R. B., Diego J. M., Martínez-González E., Lasenby A. N., 2002, *MNRAS*, 336, 1057  
 Hobson M. P., McLachlan C., 2003, *MNRAS*, 338, 765  
 Lanz L. F., Herranz D., Sanz J. L., González-Nuevo J., López-Caniego M., 2010, *MNRAS*, 403, 2120  
 Leach S. M. et al., 2008, *A&A*, 491, 597  
 López-Caniego M., González-Nuevo J., Herranz D., Massardi M., Sanz J. L., De Zotti G., Toffolatti L., Argüeso F., 2007, *ApJS*, 170, 108  
 Mahony E., Ekers R., Massardi M., Murphy T., Sadler E., Sadler 2010, *Proc. IAU Symp. 267, The Australia Telescope 20 GHz (AT20G) Survey*. Cambridge Univ. Press, Cambridge, p. 264  
 Massardi M. et al., 2008, *MNRAS*, 384, 775  
 Nailong W., 1992, in Worrall D. M., Biemesderfer C., Barnes J., eds, *ASP Conf. Ser. Vol. 25. Astronomical Data Analysis Software and Systems I. Using a Matched Filter to Improve SNR of Radio Maps*. Astron. Soc. Pac., San Francisco, p. 291  
 Nolte M. R. et al., 2009, *ApJS*, 180, 296  
 Ricci R. et al., 2004, *MNRAS*, 354, 305  
 Rider P. R., 1957, *Ann. Inst. Stat. Math.*, 9, 215  
 Sanz J. L., Herranz D., López-Caniego M., Argüeso F., 2006, in *Proceedings of the 14th European Signal Processing Conference (2006). EUSIPCO 2006 Conference, Wavelets on the Sphere. Application to the Detection Problem*. EURASIP, Darmstadt, p. 1  
 Tauber J. A., 2005, in Lasenby A. N., Wilkinson A., eds, *IAU Symp. Vol. 201. New Cosmological Data and the Values of the Fundamental Parameters. The Planck Mission*. Astron. Soc. Pac., San Francisco, p. 86  
 Taylor A. C., Grainge K., Jones M. E., Pooley G. G., Saunders R. D. E., Waldram E. M., 2001, *MNRAS*, 327, L1  
 Toffolatti L., Argüeso Gomez F., de Zotti G., Mazzei P., Franceschini A., Danese L., Burigana C., 1998, *MNRAS*, 297, 117  
 Waldram E. M., Pooley G. G., Grainge K. J. B., Jones M. E., Saunders R. D. E., Scott P. F., Taylor A. C., 2003, *MNRAS*, 342, 915

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.