Reply to Referee Report

May 21, 2012

We thank the referee for the careful review of the text. We discuss below the changes we have made incorporating your feedback. To make the changes more visible, in the new draft we have made bold and colored red the parts that are new or have changed.

> Major issues :
>
> The authors use a uniform disk model for fitting the extended sources. However many sources show a complex morphology (for example, not radial symmetric) They should quantify how the results they are presenting depend on the actual shape of the sources.

In the new draft, Section 2 has new text defining TS. Section 3 has been renamed (Validation of the TS Distribution) and has been divided into three sections. The original section is now section 3.1 (now called Point-like Source Simulations Over a Uniform Background). I will describe section 3.2 in reply to the next question.

There is also a new section 3.3 (Extended Source Simulations Over a Structured Background). In this section, we address your question by performing a dedicated Monte Carlo study of W44 simulated with a ring-type spatial model. By fitting the source with different spatial models, we assess this bias and find it to be very small.

This new section requires two new figures (Figure 6 and Figure 7). Also, in section 2.4 (Comparing Source Size) we add a line of text referring to this new section. Finally, there is a new sentence in the introduction of the paper which references the new text.

> Most of the new extended sources found in this work lie on the galactic plane at very low latitudes. Why the simulations used to validate the analysis (par. 3) do not include the diffuse Galactic background? Can the structures of this diffuse background affect the analysis results?

In section 3.2 (Point-like Source Simulations Over a Structured Background), we have added new text describing a follow up Monte Carlo study performed in the presence of the galactic plane. In addition, this new text required updating Table 1 (Monte Carlo Spectral Parameters) to include the new simulation parameters. There is also a new figure (Figure 5) which shows the distribution of TS_{ext} for these new simulations.

> An important result of this work is that some 2FGL sources could be fake due to the residuals induced by modeling the extended sources as point-like. It may be worth having a table that lists the sources that are not longer needed in the modeling of the sky (alternatively this information could be included in table 4). They must also be mentioned in the captions of figures 16, 17, 19, 21 and 22.

We have added a new table (Table 6 in the current draft, named Nearby Residual-induced Sources) which contains this information. There is also new text in section 9 (Extended Source Search Method) which refer to this table. We have also updated the requested captions to list names of 2FGL sources that were removed.

> Minor issues :
>
> Par. 3, eq 8 : The authors should explain the symbols used

The symbols are now described after the equation.

Par. 7, row 386 : W30 is cited 2 times

Fixed

Par. 10.3, rows 650-651 : The authors should explain the relation between r_{68} and σ

The relation between r_{68} and σ was already included in the text in section 2.4 (Comparing Source Sizes). But we restructured the text in that section to bring to prominence the relevant equations.

Par. 10.4, row 674 : The authors should quantify the statistical improvement using one single extended source instead of 2 pointlike

The numbers are now included in the text

Par. 10.6, row 720 : The authors should explain better how you distinguish if a source is "physically separate" or not

The text has been clarified

Par. 10.8, row 768 : The authors should quantify the statistical improvement using one single extended source instead of 2 pointlike

The numbers are included in the text

Fig. 2 : Is the y axis logarithmic ? Please clarify.

The legend now says "log(PDF) [Arbitrary Units]"

Fig. 14 : The authors should specify the spectral index of the two plotted power-laws (or else remove them).

The spectral indices have been added to the caption.

++++++++++++++++ Report on statistical methodology

Search for Spatially Extended Fermi-LAT Sources Using Two Years of Data

J. Lande et al. (ApJ86244)

a) line 98 states: In the pointlike package, "The data are binned spatially, using a HEALPix pixelization, and spectrally (G orski et al. 2005) and the likelihood is maximized over all bins in a region." Why is any binning used? Maximum likelihood estimation (MLE) is best used on exact photon positions, and binning is not required (as it is with older chi-square-minimization techniques). Binning is ill-advised as it results in a loss of information, and sometimes in a bias. As this assumption is built into pointlike, perhaps it cannot be changed by the authors for their analysis.

The Fermi LAT analysis package gtlike has an unbinned likelihood mode which is indeed slightly more precise. But running an unbinned analysis is very slow because Fermi collects a large number of photons, many of which do not contribute very much to the detection of a source. gtlike in binned mode is much faster and is therefore the most common tool for LAT analysis. Because the LAT collaboration has extensively tested binned gtlike for spectral analysis, we do not expect it to introduce any large bias in our spectral analysis result.

But even binned gtlike is prohibitively slow for the extended source spatial fitting described in this paper. That is why we developed this new functionality inside of pointlike. Even though pointlike performs a fully binned analysis, it used the healpix hierarchy for spatial bins and scales the spatial bin size with energy. This ensures that the spatial binning is always smaller than the inherent spatial resolution of the instrument. The validations of this method presented in this paper allow us to be confident in our results.

b) The likelihood ratio test (LRT, eqns 1 & 9) is a classic technique of MLE that has been largely replaced by statisticians (since the 1980s) by closely related measures that take into account the complexity of models and size of the dataset. The LRT has serious deficiencies: it tends to accept the more complex model as sample size (i.e., source counts) increases (or equivalently, it assesses only bias and not variability), and it is inapplicable when parameters (e.g. source sizes) are near zero (discussed in the ApJ by Protassov et al. 2002). I also am not confident that the chi-squared approximation in eqn 8 is correct (please refer to a statistics book not an old ApJ article, or consult a statistician).

Model selection today is instead guided by 'penalized' measures, and the most widely used is the Bayesian Information Criterion (BIC,`http://en.wikipedia.org/wiki/Bayesian_information_criterion`). It is very easy to compute, essentially an offset from the LRT scaled to the log of the counts involved. The issues are discussed in detail in the volume "Model Selection and Multi-Model Inference" (K.P. Burnham & D. Anderson, 2nd ed, 2002); see also lectures (especially #14) at `http://myweb.uiowa.edu/cavaaugh/ms_seminar.html`.

I guess you could argue that your Monte Carlo experiments (Figs 3-4) take the flux and degrees of freedom into account for the LRT statistic. But if you want your method to be widely used, it should not rest on a weak and obsolete statistical foundation. Altogether, I recommend moving to the BIC. Your approach of Monte Carlo testing of the significance of LRT values should also work for BIC values.

Concerning the extension significance test, it actually relies not exactly on Wilk's theorem (which does not hold for the extension test) but on Chernoff's theorem (see "On the Distribution of the Likelihood Ratio", `http://www.jstor.org/stable/10.2307/2236839`). we also extensively validated it against Monte Carlo simulations. Furthermore, the high energy astrophysics community has a lot of experience with the likelihood ratio test applied to nested models. So we are very confident that we can apply it to the actual data.

Concerning the test comparing two point-like sources to one extended source, we find the desire for a Bayesian test more compelling since the models are not nested. We considered the BIC test but found that it would not work well for LAT data. The reason for this is described in the text. We alternatively considered the AIC test and found that it is mathematically equivalent to the test we actually performed, putting the test on firmer ground. We have added a new paragraph to Section 5 (Testing Against Source Confusion) that describes this.

Finally, I can suggest that the lead author discuss the issue with experts at his institution, as Stanford's Dept. of Statistics is universally recognized as the best in the nation. They might provide a balanced judgment whether the BIC is substantially better than the LRT in the situation under study.

Thanks for the suggestion. In addition to the above changes, we have improved the discussion on the algorithm for fitting two point-like sources at the same time. For clarification, we have added a few minor grammatical or stylistic improvements to the text. Also, we have also fixed a few bibliographical entries.