We thank the referee for their very useful comments on the paper. We have addressed their concerns which have included adding additional figures and text (shown in red) to the paper. Our response to each point can be found below.

Reviewer's Comments:

Summary and aim of the paper:

This paper acknowledges the importance of the confusion noise and introduces a technique to

overcome the confusion noise in SPIRE images. This is relevant to the current archival investigations to dig deeper into the Herschel FIR images and the technique could also be used, potentially, in other settings where crowdedness is an issue, such as observations of the galactic center or Andromeda.

1) It is not clear how the positional priors are used. Is it based on 24 micron images or 100 micron?

sounds like 100 micron is used for simulated data and 24 micron is used for the obs data. How do you justify this? the treatments should be exactly the same for both simulation and obs data.

In the test on simulations, we use 100 micron fluxes as a prior and 24 micron fluxes for the observational run. Ideally we would also use 24 microns fluxes in our simulation test, however the Lacey et al. simulation does not include 24 micron fluxes and so the next best thing is 100 microns. We note that we use a 50 micro Jansky cut off, the typical flux cut used for 24 micron based prior source extraction. This will give us more sources than if we were able to perform a cut at 24 microns.

To show how XID+ performs for a different prior list and against a competing code, we have run XID+ using the prior cut used in Wright et al. 2016 (doi: 10.1093/mnras/stw832) when testing the LAMBDAR source extraction code.

This prior cut uses sources which have an r band magnitude brighter than 19.8 or which have an 250 micron flux greater than 4 times the noise (resembling the sources that would be picked up blind source extraction).

2) What if you use the Hubble resolution as your positional prior?

Your choice needs to be clearly justified and stated.

You should justify how accurate the MIPS 24 micron positions are with respect to HST.

We assume the positional information on the prior list is correct. Le Floch et al. 2009, found 24 micron positions to be accurate to about 2’’ in respect to K band catalogues. This corresponds to a ninth of the FWHM of the SPIRE beam at 250 microns, thereby making this a reasonable assumption. We have added a footnote regarding this assumption when we talk about the positional information in our model, being treated as deterministic.

3) How do you use the positions from the 24 micron in a coarser pixels of the SPIRE.

Do you use fractional SPIRE pixel information on the source position?

How is the convolution with the PSF done? In order to use higher accuracy positional information in a lower resolution

image, I suggest you to convolve at higher resolution and deconvolve back to the SPIRE resolution. please exactly write what you are doing.

We take the beam of SPIRE at 1’’ second resolution and convolve with the positional information from the prior catalogue source position. We have provided a brief description in the text after equation 2.

4) How are you putting flux priors based on 24 micron maps for the SPIRE maps? or is it 100 micron PACS image?

Assume a FIR template? formula? This needs to be clearly stated and elaborated.

I don’t see anywhere you talk about the flux priors. is it a flat top hat prior? gaussian? how large in dex are your flux priors.

We talk about flux priors in the Discussion. It is uniform with, upper bound of 1000mJy and lower bound of 0.01mJy, however we have added an additional sentence in section 2.1 after equation 5 to clarify our flux prior.

5) How deep can you probe with your technique below the confusion noise limit? indicate the number in mJy.

you can define a criteria for detection, and see how faint you can probe to claim detection.

As we get the full posterior out, the posterior could be used for the very faintest sources. However, when the posterior begins to approximate a Gaussian distribution, this indicates the data has enough power to constrain the model to what is typically described as a detection. We have added a section in the paper and an additional figure to describe this in more detail.

6) Do stacking on your sources at SPIRE bands for different classes of galaxies (cut based on mass, redshift or SFR),

and compare it with what you get from the median or mean of the same sample where you have individual detections for.

This is a crucial thing to do. Doing stacking on the simulated images is easy and this is a big motivation to try to convince people that it is worth to detect individual sources.

We appreciate the concern from the referee about demonstrating how faint our method can go. Following your suggestion about stacking, we have divided simulated catalogue by Stellar Mass and redshift and carried out a naïve stack on each bin. For sources that have a flux greater than ~ 4mJy, XID+ performs better than stacking. We note that this 4mJy level is about the level the posterior of sources begin to tend to a Gaussian (as described in response to question 5)

7) Where is Appendix A? you say you describe Stan in more details there, but there is no Appendix A.

This was a latex error which now sorted. Our Stan probabilistic model is now included as an appendix.

8) What is the dimensionality of your problem for each segmented map by HEALPix? How many sources on average are you fitting in one segmented map?

For a HEALPix tile at order 9, we are fitting over 600 sources on average to ~10,500 pixels at 250 microns, ~ 5500 pixels at 350 microns and ~ 2500 pixels at 500 microns. That means we are fitting over 1800 parameters simultaneously since we fitting the three SPIRE fluxes for each source and a global background and confusion noise for each SPIRE band. We have added a paragraph to the text to give an example.

9)In Figure 2, all the lines are dashed, the middle one should be solid according to the caption.

Caption for Figure 2 has been corrected (and for Figures 3 and 4)

10) In Figure 3, what is the reason behind the under-prediction? Is it because you are over predicting the confusion noise?

can you relax the assumption that \simga\_confusion is constant across all the image and instead assuming it follows a normal distribution? would that help?Are you fitting \sigma\_conf?

We have investigated this by looking at background and confusion noise estimate and believe it is due to us ignoring the correlation associated with confusion noise. \Sigma\_conf is being fitted at the same time, and is assumed to be global across the tile. Fitting individual \sigma\_conf to each pixel would require as many parameters as there are pixels and thereby drastically increasing the computational problem.

As Stan develops more sophisticated inference algorithms, we will re-visit whether we can add the correlation component to confusion noise in latter versions of the code.

We note that DESPHOT also underestimates the flux, but to a greater extent.

11) in Figure 4: three points need to be clarified:

1) you mention at fluxes less than 25 mJy the Z score is distributed around 0 with width of 1 and I can not read this from the subplot (a) which corresponds to the

250 micron image. Sounds like the flux you should quote for that is 2 mJy not 25 mJy.

We were perhaps too optimistic in our interpretation of Figure 4, and have changed the wording in the text appropriately. The distribution is close to zero, though slightly above due to the flux being underestimated. Above 25 mJy, the distribution appears to be over 1 sigma away.

2) why do we see a drastic change in the Z score for the 500 micron image? it drops at the highest flux bin.Discuss and explain this in detail.

There are only 4 sources in the highest flux bin for 500 microns and so the drop isn’t significant.

3) Why is the z-score bad at the bright end? it means your error estimate is too narrow for bright ones. Why is that?

The probability distribution for the brightest sources is highly constrained (due to there being more pixels in which the psf can be accurately fitted). Therefore the uncertainty will be smaller than that of fainter sources. This is why brighter sources have higher z-scores despite the same flux accuracy.

12) Figure 6: Plot created with Foreman-Mackey et al. (2014). what does it mean? Needs explanation.

This refers to code, which asks to be cited as Foreman-Mackey et al. (2014), however we have updated this figure with our own plotting routine in order to address the comments below.

13) Figure 6: Show the image of the two sources to give visual impression of what is meant

to be 2” apart. you should explain in more detail what the three panels are. probably show the true fluxes with a red color to distinguish between the DESPHOT which is also shown in blue.

We have updated the figure to include the aesthetic suggestions of the referee, including an inset of the map to illustrate where the sources are.

14) Is Figure 7 showing simulated data or the actual obs data? Could you plot both and compare? Could you also fit a line to dense part of the observed 2D density and over-plot it ?

what about other regions that the model don’t go through? what is the implication?

The figure showed actual observed data, however, as suggested, we have added two additional sub-plots: one showing the true simulated colour-colour space, and another showing that returned by XID+. From this, we see the XID+ is unable to put tight constraints on the colour-colour-space, despite there being a tight correlation in the simulation. The constraints placed on the observation colour-colour space appear tighter. This could be due to the slight difference in selection (i.e. MIPS 24 micron rather than PACS 100), or it could be to less variation in intrinsic SEDs.

We appreciate the referee’s suggestion about modeling the colour-colour space. We are currently investigating adding constraints to colour-colour space by modeling it at the same time as fitting the maps. We therefore leave the modeling to that future paper.

15) Figure 8, the cumulative curves seem to plateau at a level that can not describe the observed background light at the SPIRE bands. What is your theory to explain this?

We have replaced Figure 8 with a table since on hindsight we feel the figure does not provide much information above that of a table. As an additional comparison, we have added the contribution to the CIRB from stacking 24 micron sources, as measured by Bethermin et al. 2012. This shows some of that excess background is from 24 micron sources that are too faint at the SPIRE wavelengths to be fully detected by XID+. The remaining excess background is from sources not detected at 24microns. We have added this explanation to the text.

How would a DESPHOT analysis look like on that plot? is it basically oliver et al. 2010?

As DESPHOT is fitting the same map, the contribution to the CIRB would be similar to XID+. However, DESPHOT would be wrongly associating the CIRB to the different objects for the reasons discussed previously.

16) How is Gelman et al. 2013 different from Gelman and Rubin test?

They are the same, with the exception that chains are now split in two .

17) What is the nature of those sources for which the chains are not converged (R>1.2)? can you visually inspect that? are they faint sources adjacent to a super bright source? or they are sources all faint next to each other?

The sources that have a high Rhat also have a very low number of effective samples

18) are you fitting each image separately? show the average FIR SED of galaxies

grouped in mass, redshift, SFR bin at 250,350,500 micron? do they follow the empirical templates like Chary and Elbaz 2011?

Do they follow the theoretical FIR templates from for example dusty radiative transfer simulations? This can also reveal if you’re getting sensible result from your experimentation. you should always get f\_{250}>f\_{350}>f\_{500} where f is the rest frame flux at a given band.

We fit the images at the same time, but at the moment they are treated independently. We appreciate the referee’s suggestion about fitting the FIR SED of the objects. A thorough investigation of the FIR SEDs from XID+ results is planned as part of HELP and is beyond the scope of this paper

19) fit the images simultaneously. meaning fit 250,350 and 500 micron image

together and show how the posterior changes. Show the resulting FIR SED.

We fit the images at the same time, but are treated independently. As part of the SPIRE Prior model work that we make reference to in the discussion (and above in the response about colour-colour space), we are investigating how the joint posterior distribution between the bands, changes as additional constraints to the colours are imposed and this will be the subject of a future paper.

20) If you are doing SPIRE, why don’t you also do PACS?

As part of HELP, we are applying XID+ to PACS data. For this paper we have focused on SPIRE as this data is the most value for HELP and where XID+ will be most useful.

21) how are you estimating background? is it constant over the whole image? constant for all segmented map?

We estimate the background by assuming it is constant across each segment. We note that this is an advance over DESPHOT, where the background was not fitted simultaneously with the fluxes.