

# A SIMPLE STATISTICAL APPROACH FOR STAR-GALAXY SEPARATION

Basílio Santiago and the DES-Brazil group

## 1. USING SPREAD MODEL ONLY

This text presents a simple prescription to separate stars from galaxies using spread-model values and their associated uncertainties. The prescription may be extended to include other measurable quantities.

Let a source have a spread model measurement  $s_m$  in some passband, with an associated uncertainty  $\sigma_s$ . We may ask: what is the probability that this measured value corresponds to a *true* spread-model value is  $s_t$ ,  $P(s_t|s_m, \sigma_s)$ ? If the error distribution in spread model is a Gaussian, this probability (not yet properly normalized) is  $P(s_t|s_m, \sigma_s) = \exp[-(s_m - s_t)^2/2\sigma_s^2]$ .

We may also know *a priori* the probability that a randomly selected source with true spread-model  $s_t$  is a galaxy,  $P_G(s_t)$ . This probability function may come from a sample of bright sources with well measured spread-model and with HST imaging available, so that we know which objects are stars and which ones are galaxies. Alternatively, we may consider the information we have on bright sources, for which  $\sigma_s \simeq 0 \rightarrow s_m \simeq s_t$ . We notice that bright sources form two distinct loci in  $s_m$ . Therefore, we may approximate  $P_G(s_t)$  as

$$P_G(s_t) = H(s_t - s_0), \quad (1)$$

where  $H(x)$  is the Heavyside or step function.  $H(x) = 1$  for  $x > 0$  and  $H(x) = 0$  for  $x < 0$ . In other words, at the bright end we may find some value of  $s_t = s_0$  which nicely and cleanly separates actual point sources from galaxies. For DES, one often uses  $s_0 = 0.002$  or  $s_0 = 0.003$ .

We may also write the probability that a randomly picked source with spread model  $s_t$  is a point source:

$$P_S(s_t) = 1 - P_G(s_t) = 1 - H(s_t - s_0) = H(s_0 - s_t), \quad (2)$$

Assuming that we know the probability distribution  $P_G(s_t)$ , we may then compute the probability that a source with measured  $s_m \pm \sigma_s$  is a galaxy with true spread model  $s_t$ . This is simply

$$P(G, s_t|s_m, \sigma_s) = P_G(s_t) \exp[-(s_m - s_t)^2/2\sigma_s^2]$$

If we now integrate over all  $s_t$  values, we simply have the probability that the source is a galaxy, given the observed spread model values and uncertainty:

$$P(G|s_m, \sigma_s) = \frac{\int_{-\infty}^{\infty} P_G(s_t) \exp[-(s_m - s_t)^2/2\sigma_s^2] ds_t}{\int_{-\infty}^{\infty} \exp[-(s_m - s_t)^2/2\sigma_s^2] ds_t}, \quad (3)$$

Using the prior knowledge about  $P_G(s_t)$  from sources at the bright end and approximating it by a step function as in eq. (1) we then have

$$P(G|s_m, \sigma_s) = \frac{\int_{s_0}^{\infty} \exp[-(s_m - s_t)^2/2\sigma_s^2] ds_t}{\int_{-\infty}^{\infty} \exp[-(s_m - s_t)^2/2\sigma_s^2] ds_t}, \quad (4)$$

After some algebra we can prove that

$$P(G|s_m, \sigma_s) = \frac{1 + \text{erf}[(s_m - s_0)/(\sqrt{2}\sigma_s)]}{2}, \quad (5)$$

where  $\text{erf}(x)$  is the error function. Notice that if

$$(s_m - s_0)/\sigma_s \gg 1 \rightarrow \text{erf}[(s_m - s_0)/(\sqrt{2}\sigma_s)] \simeq 1 \rightarrow P(G|s_m, \sigma_s) = 1 \rightarrow P(S|s_m, \sigma_s) = 0$$

as expected. In the other extreme, if

$$(s_m - s_0)/\sigma_s \ll -1 \rightarrow \text{erf}[(s_m - s_0)/(\sqrt{2}\sigma_s)] \simeq -1 \rightarrow P(G|s_m, \sigma_s) = 0 = 1 - P(S|s_m, \sigma_s)$$

Again, this is just as expected, since spread-model values much smaller than  $s_0$  are expected to occur for point sources.

Also, if  $s_m = s_0$ , then  $\text{erf}[(s_m - s_0)/(\sqrt{2}\sigma_s)] = \text{erf}(0) = 0 \rightarrow P(G|s_m, \sigma_s) = P(S|s_m, \sigma_s) = 0.5$ . This last condition is just expected once more. By design, we use  $s_0$  as a discriminator between what is a galaxy and what is not. So, at this limiting value, a source is not clearly defined as either a galaxy or a point source.

Thus, if one wants to tag objects either as a galaxy or as point source based on what is the larger probability, eq. 5 above is exactly the same as simply cutting the sample at  $s_m = s_0$ : if  $s_m > s_0 \rightarrow$  galaxy; if  $s_m < s_0 \rightarrow$  point source.

But one may use eq. 5 to determine a probability of each source being a galaxy or a point source, carrying out this probability throughout any future analysis. Alternatively, one may also define a galaxy sample using some other criterion than simply  $P(G|s_m, \sigma_s) > P(S|s_m, \sigma_s)$ . For instance, one may draw a *more reliable* galaxy sample using a criterion such as

$$P(G|s_m, \sigma_s) = 2P(S|s_m, \sigma_s) = 0.666$$

This means that  $\text{erf}[(s_m - s_0)/(\sqrt{2}\sigma_s)] = 0.332$ . Looking at a table for  $\text{erf}(x)$ , this corresponds to  $x = 0.31 \rightarrow (s_m - s_0)/(\sigma_s) = 0.438 \rightarrow s_m = 0.438\sigma_s + s_0$ .

In other words, one may optimally select a sample of galaxies based on *a cut in the probability of objects being galaxies rather than on a cut in spread-model* or on a combination of parameters. One may also estimate what is the probability cut-off of any proposed prescription for selecting galaxies. For instance, suppose we select our galaxies by using  $s_m > s_0 - 3\sigma_s \rightarrow (s_m - s_0)/(\sigma_s) > -3$ , as has been recently proposed. In this case, eq. 5 gives

$$P(G|s_m, \sigma_s) = \frac{1 + \text{erf}(-2.12)}{2} \simeq 0$$

Such a cut-off will certainly increase completeness of the galaxy sample, but at the expense of essentially bringing any source, however small its probability of being a galaxy, into this

sample. This means a high risk of sacrificing the sample purity. Of course, if one is sure that at a certain magnitude (or S/N) level, galaxies by far dominate the whole sample, purity may not be a serious issue. This may, in fact, be the case at some very faint flux level.

## 2. INCORPORATING APPARENT MAGNITUDES

We may in fact incorporate apparent magnitude as an extra parameter in estimating the probability that some source is a galaxy. We may do it in a very similar way as we did for spread model, namely by using something analogous to eq. 3

$$P(G|m_m, \sigma_m) = \frac{\int_{-\infty}^{\infty} P_G(m_t) \exp[-(m_m - m_t)^2/2\sigma_m^2] dm_t}{\int_{-\infty}^{\infty} \exp[-(m_m - m_t)^2/2\sigma_m^2] dm_t}, \quad (6)$$

where  $m_m \pm \sigma_m$  are some magnitude measurement and its associated error,  $m_t$  is the *true* (error free) magnitude in the same filter, and  $P_G(m)$  is the probability that some randomly picked source of magnitude  $m$  is a galaxy. This may be estimated *a priori* by:

$$P_G(m, l, b) = \frac{N_G(m)}{N_G(m) + N_S(m, l, b)} = 1 - P_S(m, l, b), \quad (7)$$

where  $N_G(m)$  and  $N_S(m, l, b)$  are the expected mean number counts of galaxies and point sources per unit solid angle. For galaxies, these counts are available from deep photometry in reasonably large areas taken with similar filters. For stars, it is perhaps best to use a Galactic model. Notice that since stars dominate the counts of point sources and their distribution varies as a function of Galactic coordinates, the probability distribution of finding a galaxy as a function of magnitude depends on direction on the sky. The distribution of QSOs may also be taken from deep surveys.

Considering that  $P_G(m, l, b)$  as given by eq. 7 is not even an analytical expression, eq. 6 cannot be easily integrated to yield  $P(G|m, \sigma_m, l, b)$ . But it can be computed numerically given the curves  $N_G(m)$  and  $N_S(m, l, b)$ .

With apparent magnitudes incorporated, a new estimate of the probability that a source is a galaxy would be

$$P(G|s_m, \sigma_s, m, \sigma_m, l, b) \propto P(G|s_m, \sigma_s)P(G|m, \sigma_m, l, b), \quad (8)$$

where the first factor in eq. 8 above is given by eq. 5 and the second factor is given by eq. 6 with integrand given by eq. 7. Expression 8 is stating that the effect of spread model on the probability of an object being a galaxy is independent from the effect of the source's brightness.

One has to be careful in dealing with the normalization of this new probability though. The reason is that we can write (see eqs. 2, 3 and 4)

$$P(S|s_m, \sigma_s) + P(G|s_m, \sigma_s) = 1$$

and (see eq. 6 and 7),

$$P(S|m, \sigma_m) + P(G|m, \sigma_m) = 1$$

In other words, both factors add to one individually. To bypass this, we may compute  $P(G|s_m, \sigma_s, m, \sigma_m, l, b)$  as an equality in eq. 8, then compute  $P(S|s_m, \sigma_s, m, \sigma_m, l, b) = P(S|s_m, \sigma_s)P(S|m, \sigma_m, l, b)$  in an identical way and renormalize both at the end, so that they add up to 1.

### 3. COMPUTING THE PROBABILITY ESTIMATES

The DES portal QA Coadd tool is likely the ideal place to compute these probability estimates for each source in a DES catalog. It efficiently deals with DES catalog data containing positions, spread model and magnitudes and their uncertainties. Its modules already incorporate model number counts for stars (from AddStar on a tile by tile basis) and for galaxies (in fact, these curves are available in Quick Reduce and were taken from Capak et al (2007, ApJS, 172, 99), but can be easily transferred to QA Coadd). Thus, all the necessary quantities to compute  $P(G|s_m, \sigma_s, m, \sigma_m, l, b)$  and  $P(S|s_m, \sigma_s, m, \sigma_m, l, b)$  using the prescription above are available.

As a final remark, colour information can be also factored into these probabilities. One simple way to do it is to use galaxy and stellar number counts in the  $(g - r)$  vs.  $(r - i)$  plane. We can use stellar population synthesis from AddStar and observed galaxy counts in colour-colour space to determine the prior probability that some randomly selected source is a galaxy or a star, given its colours. The probability that the source is a galaxy based on a Gaussian distribution of measured colour uncertainties and on the prior probabilities over colour space should then be easily determined, in an analogous way to what was done with magnitudes. The colour based probability may then be combined with those based on spread model and magnitudes and the final probabilities using all parameters can then be renormalized to unity as described in the end of the previous section.

### 4. INITIAL IMPLEMENTATION OF THE METHOD

We have implemented the algorithms described above using Python. We used the S/G challenge test file, containing about 71k sources from the DECam COSMOS field in about a 1 sq deg region. See details at the RedMine wiki page

[https://cdvs.fnal.gov/redmine/projects/des-sci-verification/wiki/SG\\_separation\\_challenge\\_details](https://cdvs.fnal.gov/redmine/projects/des-sci-verification/wiki/SG_separation_challenge_details)

The prior assumptions about the distribution of stars as a function of magnitude and colours were taken from AddStar runs of tiles in and around the COSMOS field. The galaxy distribution as a function of magnitudes was assumed to be that from Capak et al (2007, ApJS, 172, 99). For the galaxy distribution in colour-colour space ((g-r) vs. (r-i)) we assumed, for simplicity, that it is uniform over the range covered by the stellar locus.

The code sgsep.py computes probabilities of a source being either a galaxy or a point source based on spread model alone, on the magnitude alone and on the colours alone. The 6 probability values are then output to an ascii file.

Figure 1 shows all sources in the S/G Challenge test file in the *spread\_model\_i vs. mag\_auto\_i* plane. The points are colour coded according to the probability of a source being a galaxy based only on spread model,  $p_g(s_m)$ , as given by eq. 5 and assuming  $s_0 = 0.002$ . The colours are as follows: cyan:  $p_g < 0.2$ ; blue:  $0.2 < p_g < 0.4$ ; green:  $0.4 < p_g < 0.6$ ; red:  $0.6 < p_g < 0.8$ ; black:  $p_g > 0.8$ .

There is a clear stratification where higher  $s_m$  values correspond to higher probabilities of the source to be a galaxy. The separation is very clear cut at bright magnitudes (i.e., either very small or very large  $p_g(s_m)$  values) because of the very small uncertainties in spread model at high S/N levels. At fainter magnitudes the S/G separation is not as simple, with many objects with  $0.2 < p_g(s_m) < 0.8$ , attesting the increasing difficulty in separating sources and the larger uncertainties in the measured values of spread model.

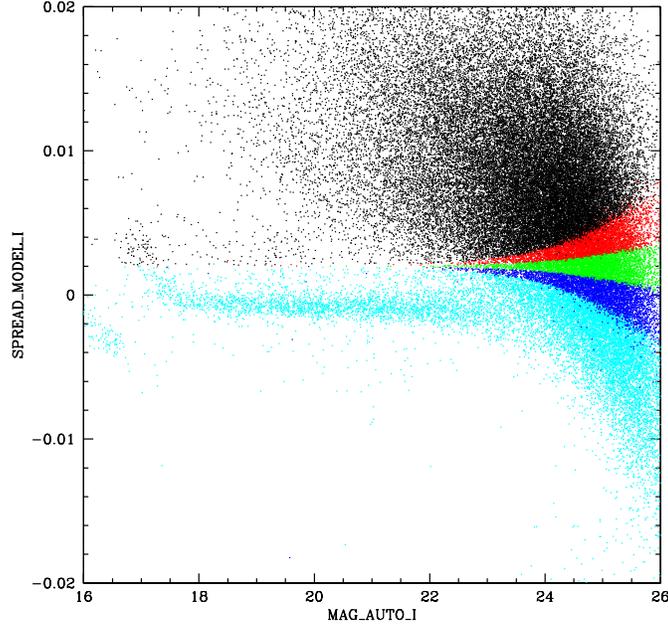


Figure 1: Spread model ( $s_m$ ) plotted against magnitude, both in  $i$ -band. The different colours correspond to different probabilities of the source being a galaxy. Cyan:  $p_g < 0.2$ ; blue:  $0.2 < p_g < 0.4$ ; green:  $0.4 < p_g < 0.6$ ; red:  $0.6 < p_g < 0.8$ ; black:  $p_g > 0.8$

Figure 2 shows the same plot as in Figure 1, but sources are colour coded according to the galaxy probabilities exclusively based on the measured magnitude and its uncertainties, and on the assumed priors for the magnitude distributions,  $N_G(m)$  and  $N_S(m, l, b)$ , as given by eq. 7 above. We again used the  $i$  band in this experiment, i.e. we set  $m = i$ . The colour code is the same as in the previous figure. Again, the stratification is clear, showing that the algorithm has been correctly implemented. Higher  $p_g(i)$  probabilities are assigned to fainter sources, since galaxies increasingly dominate the source number counts as a function of magnitude.

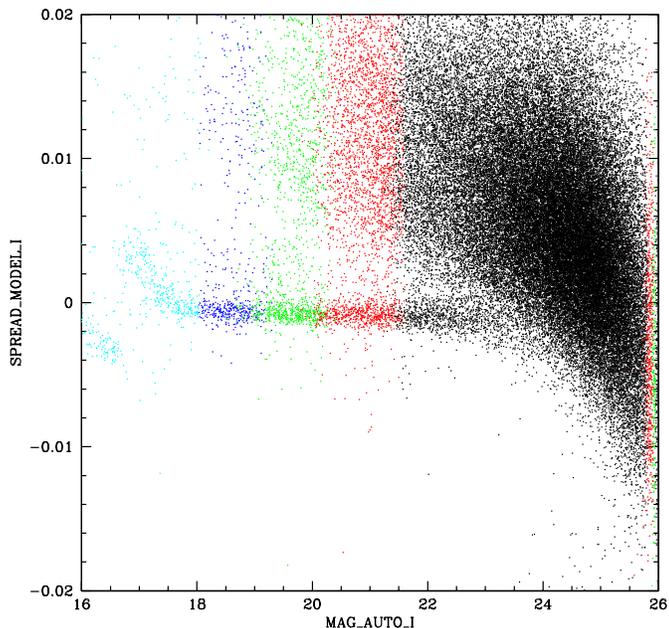


Figure 2: Same as in Figure 1, but now the sources are colour coded according to their magnitude based probabilities,  $p_g(i)$ .

As a final check to our algorithm, in Figure 3 we show the sources from the S/G Challenge test file distributed on the  $(g - r).vs.(r - i)$  plane. They are colour coded by the galaxy probabilities based on the colours alone,  $p_g(g - r, r - i)$ . As mentioned earlier, the prior distribution of stars on this plane was built using AddStar simulations made on the same region of the sky where the data come from. The galaxies were assumed to be uniformly distributed over the colour-colour plane. Based on these distributions we compute prior probabilities of a source being a galaxy or a star, analogously to equation (7) above. In fact, the algorithm to compute  $p_g(g - r, r - i)$  and  $p_s(g - r, r - i)$  is formally identical to the one used for the magnitudes and which was presented earlier. The basic difference is that we now integrate the Gaussian terms over a 2D (colour-colour) plane rather than over a single (magnitude) axis, as we did in equation (6).

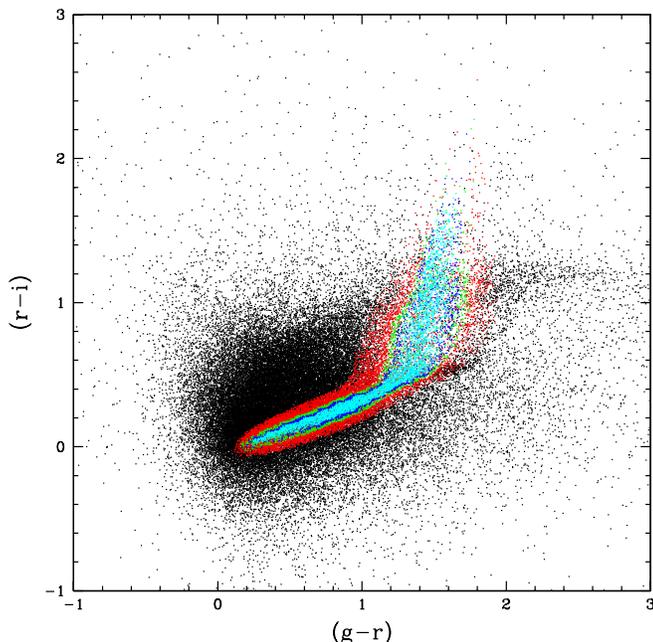


Figure 3: Sources on the  $(g-r)$  vs;  $(r-i)$  plane colour coded according to their colour-based probability of being a galaxy,  $p_g(g-r, r-i)$ .

Figure 3 again shows that our code was correctly implemented. The  $p_g(g-r, r-i)$  values are clearly smaller in regions of the  $(g-r)$ .vs. $(r-i)$  plane that lie closer to the stellar locus, as expected.

## 5. S/G CHALLENGE, FIRST RESULTS

We now apply the probability estimates we have to the S/G Challenge test sample and measure completeness and purity curves for stars and galaxies. Having 3 estimates of the probability that each source is a galaxy,  $p_g(s_m), p_g(i), p_g(g-r, r-i)$ , we need to combine them.

Our first attempt is to use their product

$$p_g = p_g(s_m) p_g(i) p_g(g-r, r-i). \quad (9)$$

We do the same for the stars and renormalize the two probabilities so that  $p_g + p_s = 1$ . Using the parameter `mu_class_acs` described in the S/G Challenge web page, we then build Figure 4. It shows galaxy completeness plotted against galaxy purity for the entire test sample. No magnitude cut was applied to the sample prior to constructing the curves. As suggested in the same wiki page, we did not include objects with `mu_class_acs = 3` or `mu_class_acs = -999` in our completeness and purity estimates. Hence, for a given  $p_{g,lim}$  cut-off value, our galaxy completeness estimate is

$$cpt_g(p_{g,lim}) = \frac{N(p_g > p_{g,lim} \ \& \ mu\_class\_acs = 1)}{N(mu\_class\_acs = 1)}, \quad (10)$$

A similar expressions holds for the stellar completeness. As for purity we get

$$pur_g(p_{g,lim}) = \frac{N(p_g > p_{g,lim} \ \& \ mu\_class\_acs = 1)}{N(p_g > p_{g,lim} \ \& \ |mu\_class\_acs| < 3)}, \quad (11)$$

And again a similar expression is used for the stars.

The solid line in Figure 4 shows the results for our Bayesian classifier. Larger completeness of course means lower  $p_g$  cut off values, at the expense of decreasing purity. The dashed shows the same curve based on *class\_star\_i*, whereas the dotted one is the result of using simple cuts in *spread\_model\_i*. These curves follow the same definition of completeness and purity as presented above, only the parameter used and its associated cut-off values changed.

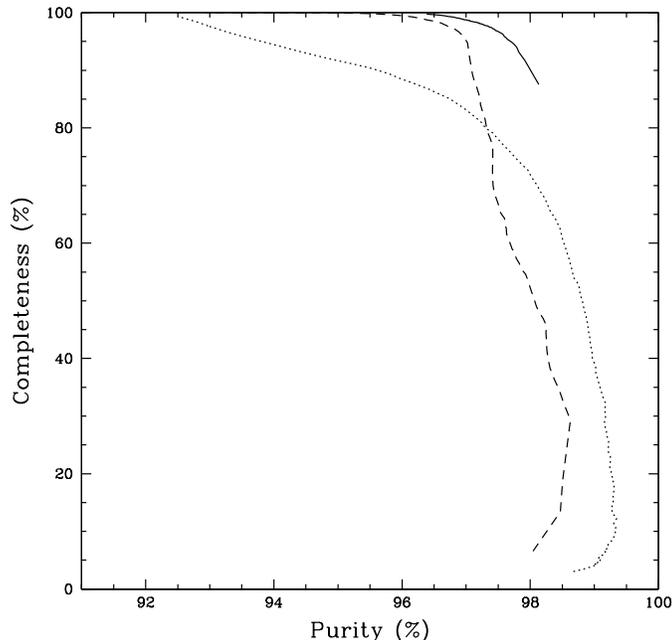


Figure 4: Galaxy completeness values plotted against purity. The solid line shows the results for our combined probability estimator. The dashed (dotted) line is based on *class\_star* (*spread\_model*). Completeness and purity were computed according to eqs. 10 and 11 in the text.

It is clear from Figure 4 that the Bayesian estimator proposed here simultaneously yields both a complete and pure galaxy sample. At 98% purity, for instance, we have a galaxy sample which is  $\simeq 91\%$  complete. If we sacrifice purity by just 1%, we increase completeness to 98%. The improvement over a simple *class\_star* or *spread\_model* criterion is visible in the figure.

In Figure 5 we show our results for the stars. The symbols are the same as in the previous figure. Again, we did not cut the sample at any S/N level. With the proposed estimator we now have much lower simultaneous completeness and purity values for the stars. Still, our estimator achieves higher simultaneous levels than the two classical S/G separators. We can

extract a  $\simeq 90\%$  pure stellar sample with a completeness level of  $\simeq 60\%$ . One caveat that affects our method is that the  $p_s(s_m)$  probability is estimated using a cut in  $s_m$  rather than  $|s_m|$  (see description of the method). This may affect the purity of the stellar sample. We are planning to improve on this probability estimator, in order to restrict it the actual stellar range in  $s$ , which is supposedly symmetric around  $s_m = 0$ .

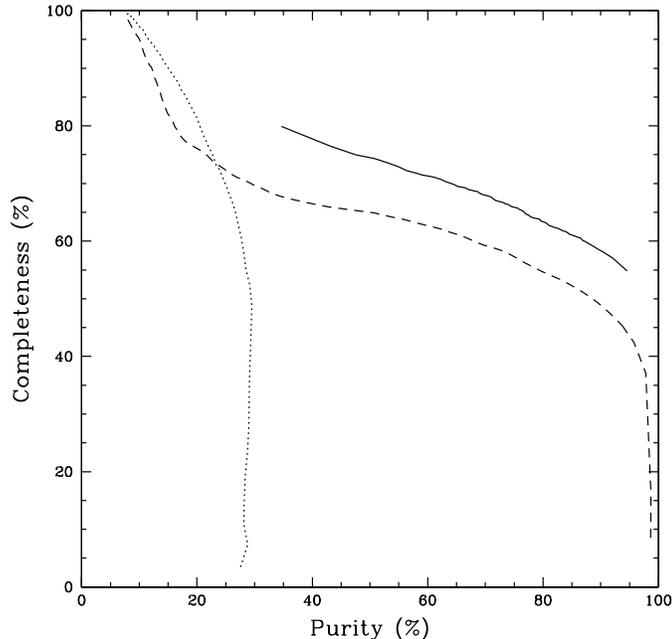


Figure 5: Same as in Figure 4, but now showing results for the stars.

Since the results for the stars are not as good as for galaxies, we have also explored other combinations of the stellar probabilities based on spread-model, magnitude and col-col space than the product as given by eq. 9 applied to stars. In Figure 6 we show these results. We tried all combinations involving  $p_s(s_m)$ ,  $p_s(i)$ , and  $p_s(g - r, r - i)$ . In all cases  $p_g$  was consistently redefined so as to make use of the same estimator as the used for the stars. Unfortunately, considering the 3 stellar probability terms individually or in pairs does not improve our capacity to draw a more complete and purer stellar sample than using the full combination of all 3 terms. Figure 6 also shows that use of the colour information is producing an artificial discontinuity in the completeness and purity values and is adding little extra information to help separate stars from galaxies.

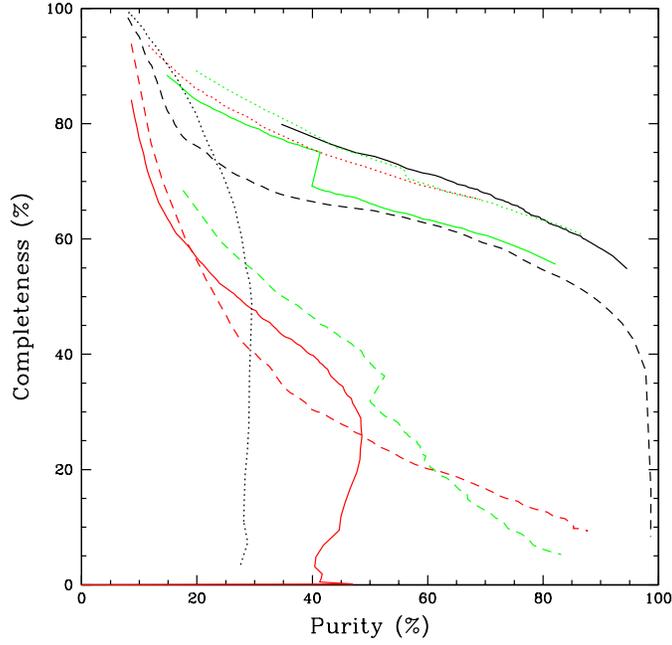


Figure 6: Same as in Figure 5, but now showing results for stellar probabilities based on different combinations of  $p_s(s_m)$ ,  $p_s(i)$ , and  $p_s(g-r, r-i)$ , as follows: solid black line:  $p_s = p_s(s_m) p_s(i) p_s(g-r, r-i)$  (same as in previous Figure); solid green line:  $p_s = p_s(s_m) p_s(g-r, r-i)$ ; dotted green line:  $p_s = p_s(s_m) p_s(i)$ ; dashed green line:  $p_s = p_s(i) p_s(g-r, r-i)$ ; solid red line:  $p_s = p_s(g-r, r-i)$ ; dotted red line:  $p_s = p_s(s_m)$ ; dashed red line:  $p_s = p_s(i)$ . In all 7 cases, the stellar probabilities were re-normalized using the condition  $p_s + p_g = 1$ , where  $p_g$  was also re-computed using the same combination of probability terms as the stars. The curves based on class star and spread model are again shown here for reference.

As a final test, we cut the stellar sample at  $i \leq 23$  and rebuild the completeness vs. purity curve. This is shown in Figure 7, for our method, as well as for class star and spread model, as the curves that lie above the ones repeated from Figure 5. There is a clear improvement in the completeness and purity levels achieved for all classifiers. In particular, the method presented here now provides 95% purity at about 92% completeness for the stars brighter than  $i = 23$ .

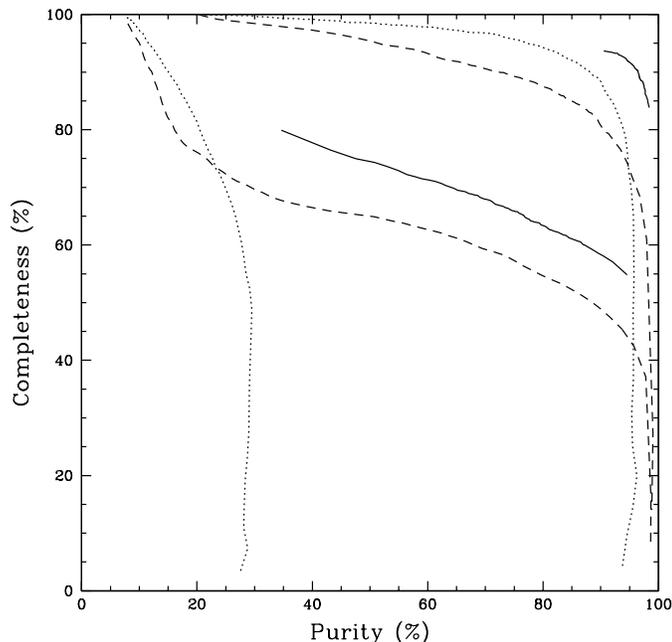


Figure 7: Stellar completeness values plotted against purity. The solid line shows the results for our combined probability estimator. The dashed (dotted) line is based on class star (spread model). The lower curves are the same as in Figure 5, where no cut in magnitude was applied. The curves lying closer to the top right of the figure correspond to sources cut at  $i = 23$ .

## 6. FINAL REMARKS, FOR NOW...

QSO probabilities may be incorporated into our method in a straightforward manner, basically requiring only that prior distributions of QSOs as a function of magnitude and in colour-colour space are provided. The probability that some source with measured spread model, magnitude and colours, along with their uncertainties, is a QSO would then be computed as

$$p_Q = p_s(s_m) p_Q(i) p_Q(g-r, r-i), \quad (12)$$

where  $p_X(i)$  and  $p_X(g-r, r-i)$  for  $X = g, s, Q$  would be computed *a priori* using expressions similar to eq. 7, but with the 3 source counts in the denominator.

And the re-normalization of the probabilities would then just be

$$p_g + p_s + p_Q = 1$$

Another improvement that can be made is to incorporate probability estimates for point sources and galaxies based on the value of class star and its associated uncertainty, in a similar way as described in section 1 for spread model. The method can also be further extended to

use extra dimensions in colour-space, magnitude space as well as spread model and class stars estimated with different filters.