
# Quasar classification with DSC. Contamination, completeness and modified priors

prepared by: Coryn Bailer-Jones, Kester Smith
approved by:
reference: GAIA-C8-TN-MPIA-CBJ-036
issue: 1
revision: 0
date: 2008-04-11
status: issued

# Abstract

We analyse the performance of the Gaia Discrete Source Classifier (DSC) to identify clean samples of quasars, as is necessary for defining the Gaia astrometric reference frame. We present overall classification performance and the trade-off between sample completeness and contamination. We further develop a method for modifying the priors of an arbitrary classifier in order to reflect different expected relative frequencies of objects, in particular the rareness of quasars compared to stars. This enables us to estimate posterior class probabilities based on an appropriate prior and not, for example, one implicity defined by the relative frequency of objects in the training data set. We apply this to a simulated situation in which quasars are 1000 times rarer than stars. At G=18.5, we can use DSC to build a quasar sample with a completeness of 51% which has a contamination of less than 0.01%, with classifications based on simulated end-of-mission BP/RP data only. Including astrometry (parallax and 1D proper motion) does not significantly alter this. At G=20 (without astrometry) this degrades to a completeness of 41% at the same upper limit on the contamination. Including astrometry only raises this to 43%. We discuss the relevance of the prior modification approach and the interpretation of the posterior probabilities for a probability threshold-selected sample. In an appendix we query the SDSS database to estimate the relative frequencies of stars to quasars as a function of G magnitude and Galactic latitude. At $b > 60^\circ$ and $g = 18$–$19$, the quasar/star ratio has an upper limit of 0.02. A definitive value is precluded by stellar incompleteness in SDSS.

# Document History

| Issue | Revision | Date | Author | Comment |
|---|---|---|---|---|
| 1 | 0 | 2008-04-09 | CBJ | Minor updates. First release |
| D | 0 | 2008-04-08 | CBJ | Comments from C. Tiede, added two discussion subsections (new tests, final word), removed contingency tables for the modified case |
| D | 0 | 2008-04-06 | CBJ | Added abstract, introduction, astrometry density plots, discussion of nominal case and of posterior probabilities, conclusions; corrected appendix. Checked |
| D | 0 | 2008-03-24 | CBJ | Replaced cycle 3 STVR (cycle 2b data) results with cycle 3 data results |
| D | 0 | 2008-02-27 | CBJ | Added the appendix with the SDSS quasar/star ratios |
| D | 0 | 2008-02-21 | CBJ | Rewrote and greatly expanded discussion of (model-based) priors |
| D | 0 | 2008-02-14 | CBJ | Modified discussion on class fractions following comments from C. Tiede |
| D | 0 | 2008-02-10 | CBJ | Working draft with figures and introductory text |

# Contents

# List of Figures

# 1 Introduction

The Discrete Source Classifier (DSC) is the algorithm which classifies Gaia objects into broad astrophysical categories, "single star", "galaxy", "quasar", "physical binary star" etc. It currently uses epoch-combined BP/RP data, supplemented with astrometry where available and, in common with the Gaia onboard detection strategy, only operates on point-sources (i.e. makes no use of morphological information).

One of the objectives of DSC is to identify a "clean" quasar sample which will be the basis of the Gaia astrometric reference frame. Quasars cannot, of course, be identified with 100% accuracy: there is an inevitable trade-off between achieveing completeness (maximizing the number true positives) and reducing contamination (minimizing the number of false positives). This technical note analyses the results of the application of the current DSC implementation to quantify the trade-off between these two factors. As DSC assigns class probabilities to each object observed, we can build different samples of quasars by varying the threshold for inclusion in the sample and thereby control the completeness and contamination.

It is important to realise that the probabilities (and thus assigned classes) of any classifier depend not only on the evidence in the data, but also on the prior probabilities. In this context, the prior is the class probabilities before one looks at the data. Priors are always present (we will have *some* idea of the relative frequencies of stars and quasars at G=20, for example), but sometimes just implicitly and sometimes it is difficult to know exactly what these priors are. This is a problem, because if the model priors are inappropriate (e.g. equal probability of star and quasar) this could bias our posterior probabilities and hence our class assignments. In particular, the relative frequency of classes in the training data set may influence the priors. This is sometimes called the "class imbalance problem" in the machine learning literature. We present a method for calculating these priors and for modifying the posterior probabilities to reflect different priors, e.g. for the case when quasars would be much rarer than stars. This enables us to vary priors and to build, post hoc, different classification models without having to modify the training data sets and to retrain the models. Using this we calculate the quasar sample completeness and contamination for a realistic quasar/star population. This approach is preferred to modifying the relative frequencies of classes in the data sets because, as we will demonstrate, these frequencies generally do not reflect the prior.

# 2 The classification model

## 2.1 Classification algorithm

DSC is currently implemented as a multi-class Support Vector Machine (SVM). It is a supervised learning algorithm which assigns, to each presented object, the probability that the object

is of class "single star", "physical binary star", "galaxy" or "quasar". [1] The classes are exclusive and exhaustive, so the probabilities sum to unity for each source. The algorithm is described in the CU8 cycle 3 software release documents (KS-003, KS-004, KS-005). Compared to those documents, the DSC code used here has been restructured, but the underlying SVM is still the same external library (libsvm, wrapped and implemented by C. Tiede). DSC has also been re-trained using cycle 3 data (during cycle 4), although the SVM parameters ($C$ and $\gamma$) were kept the same (no new "tuning").

## 2.2 The train and test data

The input data for the classification are either BP/RP or BP/RP+astrometry (two separate models). The model is trained on 2000 objects of each of the four classes (equal class fractions). The test data set has the same size and frequencies, but different objects. Both train and test data were drawn at random from the cycle 3 simulations (CBJ-029, RS-002) using a program written by C. Elting. We use the non-oversampled end-of-mission BP/RP spectra, which have a total of 120 pixels (Figs. 1 and 2). This includes about 20 "edge pixels" which contain essentially no signal. Indeed, classifications made excluding these are somewhat better than the results presented here.[2] When included, astrometry is presented as two additional input variables, the parallax and the proper motion (in the latter case the RMS of the two components).

All data, both in the training and testing sets, are noisy, using the photometric and astrometric noise models implemented in GOG. Extragalactic astrometry is just zero with noise added. The stellar parallaxes were assigned to be consistent with the physical parameters and apparent magnitude (CBJ-029, RS-002), and the proper motion is assigned using a simple stochastic model (see section 2.4 of CBJ-029 – a better model is described in CBJ-037). The astrometry in the test sample is shown in Fig. 3 (zoomed in Fig. 4) and the densities of each parameter are shown in Fig. 5. The parallax and especially the proper motion distribution for physical binaries is significantly different from that of single stars, with the former generally having much larger values (mode of the proper motion differs by a factor of about forty!). This is partly due to the different distributions over physical parameters in the two libraries: the single stars contain a relatively large number of OB stars, whereas the physical binaries are limited to the range 4000–8000 K effective temperature. Yet, the difference is so large that it may be an unintentional error. However, as our main purpose is to analyse the classification of quasars, this should not be a source of error.

[1] The cycle 3 DSC actually uses more classes, but we limit it to four for the sake of this experiment.

[2] The leading diagonals in the contingency tables for nominal results at G=20 BP/RP (Table 4) increase by up to 10% and the overall classifcation accuracy increases by 8%.

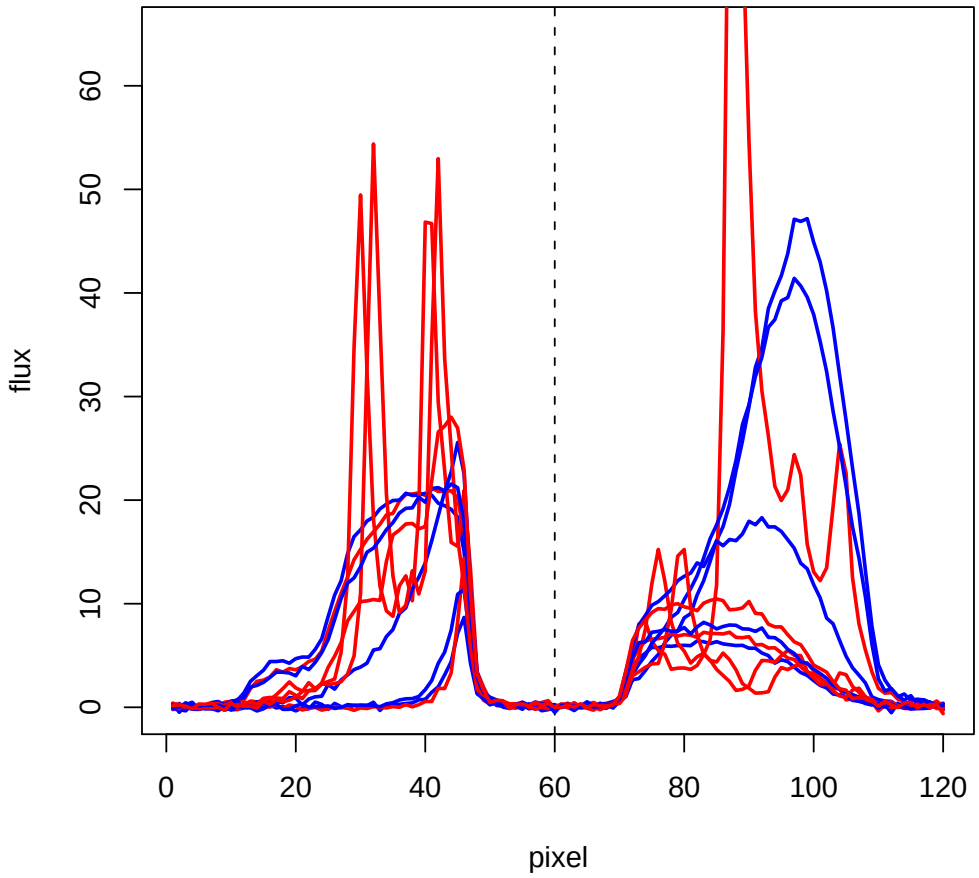


Figure 1: Random selection of five stars (blue) and five quasars (red) (G=18.5, noise included)

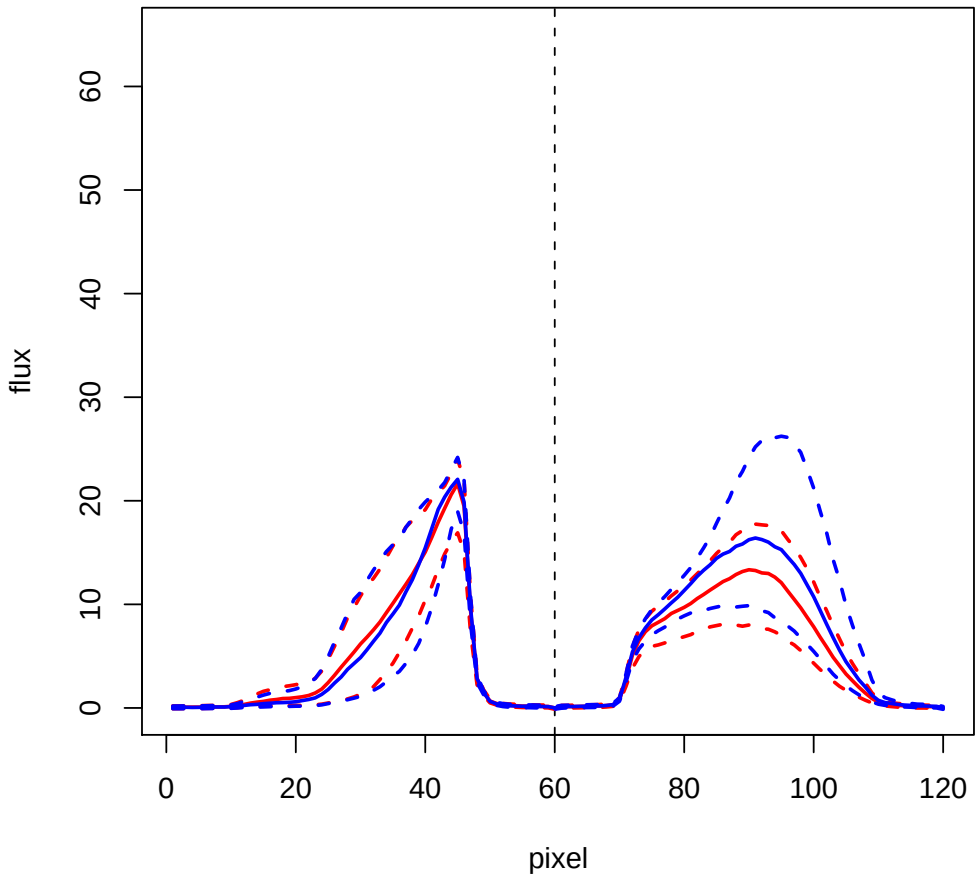


Figure 2: Median spectrum (solid line) and upper and lower quartile spectra (dashed lines) of stars (blue) and quasars (red). The upper quartile spectrum is that which defines where 25% of spectra lie above this line; below the lower quartile spectrum lies another 25% of the spectra. Equivalently, half of all spectra lie between the two quartiles. They are essentially robust versions of the $\mu \pm \sigma$ (mean plus/minus one standard deviation) spectra but allow for the fact that the flux distribution is asymmetric (e.g. cannot be negative). Spectra are for G=18.5, noise included.

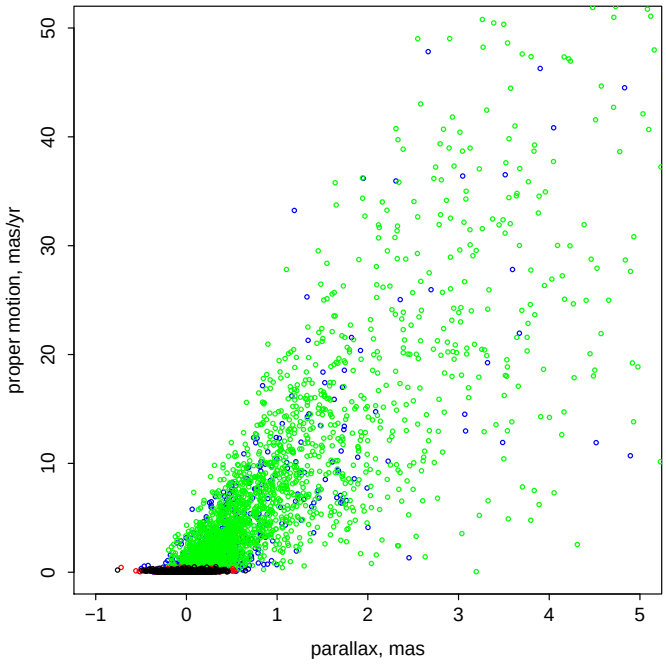


Figure 3: Proper motion and parallaxes of the 8000 objects in the test data at G=18.5 (noise included). Blue=star, green=physical binary, red=quasar, black=galaxy.

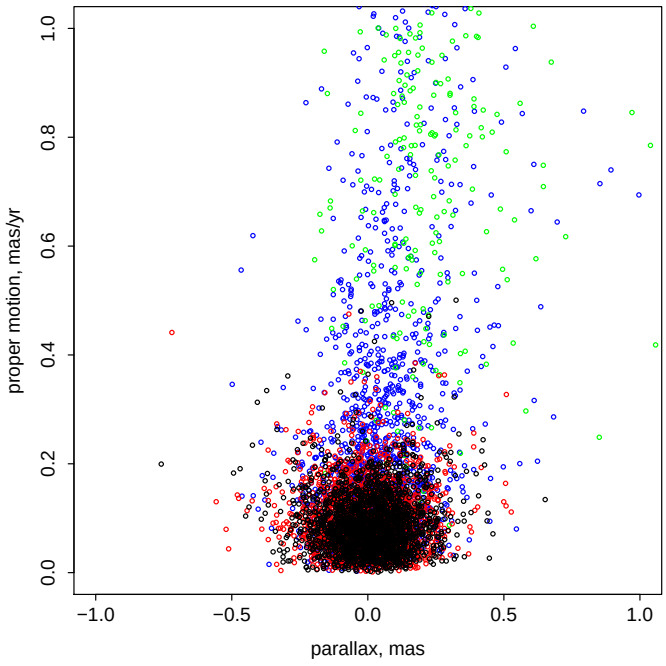


Figure 4: Zoom of Fig. 3

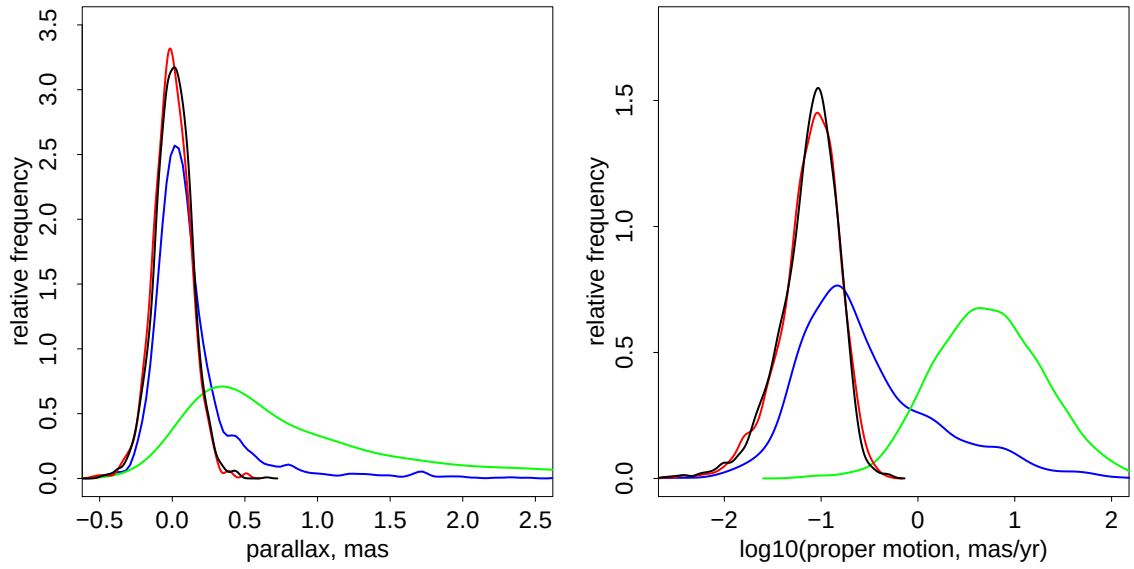


Figure 5: Distribution of the astrometry shown in Fig. 3. Note that the axis in the proper motion plot (right) is on a log scale. Blue=star, green=physical binary, red=quasar, black=galaxy.

# 3 Building samples and assessing completeness and contamination

DSC assigns to every source spectrum, $x_n$, a probability, $P(\texttt{quasar})$, that it is a quasar, and likewise for the other three classes.[3] The classifier is of course not perfect, so generally $P(\texttt{quasar}|\texttt{QUASAR}) < 1$ and $P(\texttt{quasar}|\texttt{!QUASAR}) > 0$ (the same being true for the other three classes).[4]

How should we use these probabilities? The simplest thing is to assign an object to that class for which the corresponding DSC output is the largest, "the most probable class". However, if all of the DSC output probabilities were similar, with one just sightly larger than the others, this would not be a very confident classification. It certainly could not be the basis for forming a "clean" sample of objects.

To select a sample of quasars we should decide on a probability threshold, $P_t$, such that only

[3] Strictly – and for consistency with later notation – this quantity is $P(\texttt{quasar}|x_n, \theta)$, where the class $C_j = \texttt{quasar}$, $x_n$ denotes the data for object $n$ and $\theta$ represents the DSC classifier. But for now we drop the stuff to the right of the "given that" symbol, "|".

[4] Lower case class names, e.g. `quasar`, denote estimated classes, upper case, e.g. `QUASAR`, true classes. I use the notation $P(\texttt{quasar}|\texttt{QUASAR})$ to mean "the DSC-assigned quasar probability given that the source is truely a quasar". The exclamation mark means logical "not". This notation is simply a shorthand to refer to DSC outputs for objects with known classes (e.g. in a test set). This is not to be confused with the notation $P(C_j|x_n, \theta)$, which refers to the DSC outputs in the general case.

sources with $P(\texttt{quasar}) > P_t$ are classified as quasars. If $P_t = 0$ then everything is classified as a quasar and contamination is a maximum. In contrast, in the limit as $P_t$ is increased to 1 the sample contamination is minimized but will include very few true quasars, i.e. the completeness is minimized. One of the two objectives of this technical note is to quantify how contamination and completeness vary with $P_t$ and thus identify a threshold for the quasar output which is somehow "optimal".

A *sample* for class `class` is made by selecting all objects in the test set with DSC output $P(\texttt{class}) > P_t$, where $P_t$ is user-defined independently for each class. The **completeness** of the sample is defined as the number of objects in the sample which are truely of that class divided by the total number of objects of that class in the test set. The **contamination** of the sample is defined as the number of objects in the sample which are truely of other classes divided by the total number of objects in the sample. For $\texttt{class} = j$ these are

$$\begin{aligned} completeness_j &= \frac{n_{i=j,j}}{N_i} \\ contamination_j &= \frac{\sum_{i\neq j} n_{i,j}}{\sum_i n_{i,j}} \end{aligned} \tag{1}$$

where $n_{i,j}$ is the number of objects of (true) class $i$ in the sample selected for (output) class $j$, and $N_i$ is the total number of objects of (true) class $i$ in the test set. We can also use these equations to predict the completeness and contamination for a new (unlabelled) data set, provided that this has the same class fractions as the training data (see section 4.2). It is obvious that when building samples based on thresholds, some objects may be assigned to no sample and thus remain unclassified.

# 4 Model-based class priors

## 4.1 Bayes and priors

The outputs from a trained classifier when presented with data $x_n$ are $\{P(C_j|x_n,\theta)\}$, the probability that the data is of class $C_j$ given the data and the model, $\theta$. This latter quantity reflects both the architecture of the model and the training set used to fix its internal parameters. The classifier outputs are examples of posterior probabilities in the context of Bayesian learning, and can be written using Bayes' theorem as

$$P(C_j|x_n,\theta) = \frac{P(x_n|C_j,\theta)P(C_j|\theta)}{P(x_n|\theta)} \tag{2}$$

The first term on the RHS, $P(x_n|C_j,\theta)$ is the *likelihood* of the data given the class and model. (It is an explicit function of $x_n$, although in classical likelihood analysis we think of it as a function of $C_j$.) The second term, $P(C_j|\theta)$, is the *prior probability* that, given our model, an object is of class $C_j$. These are the probabilities we would assign to an object just on the basis of our

model, before we look at the specific data on the object. This may seem like a curious quantity: how can we have any idea what the class is before we look at the data? Well, we can and we do. Imagine that I have a catalogue of all objects in the sky down to G=20 and I select one at random. I ask you "what is the probability that it is a quasar?". If you know something about the universe, you willl probably say something like "1 in 1000". You answered this without looking at the data. You probably wouldn't say "1" (definitely a quasar), so you clearly have some prior knowledge. If you had a different idea about the universe, or if I now told you my catalogue was only of the Galactic plane, you might give me a different answer, and for good reason. You might have no idea what a quasar is and, thinking that the universe consists of just stars and quasars say "0.5". Your prior will vary depending on your experience, personal bias, knowledge of the survey and model you've used. It is important to realise that *all* classification algorithms have priors. Even so-called non-Bayesian methods have priors; with these it is often just not explicit what the prior is.[5]

## 4.2 What priors are we using and how does the training data influence them?

We want to know what the assumed prior is, for two reasons. First, we would like to know what assumption our model is actually making (and not what we think it is making!). Second, we need to control the prior. We can do this in two ways, either when we train the model or by modifying the output probabilites from a trained model. We will do the latter in section 5, but first, we must calculate the priors our model is using.

In some classification algorithms and learning processes the prior is explicit. This is the case for mixture models used for classification, for example. In others algorithms, the prior is implicit, and in many of these the prior depends on the relative number of objects of each class in the training data set (the "class fractions"). Take, for example, a regression model which is trained by minimizing an error function over the entire training data set. A specific example might be a neural network trained with the residual sum of squares (RSS) error function. If we trained this on 1000 stars and just 1 galaxy, then the network will learn to recognise stars much better than galaxies (because in minimizing the error it hardly has to worry about fitting the lone galaxy). If we changed the training data (class fractions) then the model would perform differently, i.e. it would assign different probabilities on a given test set. Given this dependence we refer to the priors as "model-based priors", and the notation $P(C_j|\theta)$ reminds us that the priors depend specifically on the training data via $\theta$.

This issue of the training data class distribution influencing the model performance is well-know in the machine learning literature and is somtimes referred to as the problem of "class

[5]It is sometimes claimed that a "non-informative" prior – e.g. one in which all classes are assigned equal prior probabilities – is somehow "unbiased", but this is not true. Just take the quasar example: Can we claim that the 0.5 probability for quasars is unbiased? All priors are, in some sense, a bias. This simply reflects how we do inference: we use both data *and* prior knowledge to draw conclusions.

imbalance" of "imbalanced data sets" (e.g. Shin & Cho 2003, Visa & Ralescu 2005, Weiss 2004).[6]

But how, exactly, does the training data class distribution affect the classifications and, more specifically, the model-based priors? We might think that in the above example the ratio of the prior probabilities for stars to galaxies would be 1000 to 1, but this is not evident.[7] In the case of fitting a linear boundary between two classes to minimize the RMS error, then this may be the case, because the position of the boundary is strongly dictated by the data — the model has a very free form. But consider a logistic regression model: the shape of this is much more constrained. Once we have sufficient data to constrain its form, adding many more examples of one class will have little impact on the class boundary position and so less impact on the model-based prior. The take home message is that the model-based prior is not, in general, equal to the class fractions in the training data.

## 4.3 Calculating the model-based priors

It turns out that we can estimate the model-based priors, $P(C_j|\theta)$, directly from the trained model using a test data set. The model gives an output, $P(C_j|x_n,\theta)$, for all of the $N$ objects in the test data set for each class $C_j$. These are related to the priors via the marginalization equation

$$P(C_j|\theta) = \sum_{n=1}^{n=N} P(C_j|x_n,\theta)P(x_n|\theta) \tag{3}$$

where the sum is taken over all $N$ objects in the test data set. The second term is the probability that we draw object $x_n$ from the test data set. Clearly, this is just $1/N$. This sum is therefore just the average of the outputs for class $C_j$, and this is the model-based prior.[8]

We apply equation 3 to the three nominal data sets ($G$=15, 18.5, 20; with astrometry) to obtain the following estimates of the priors

[6]There is work in the literature which addresses this problem, typically for the case where we have very little training data for one class and want to compensate for this in the learning process to ensure that the model learns to properly recognise the minority class. Class imbalance has been demonstrated to impact neural networks and SVMs (for both classification and regression problems) and classification trees. However, we have not found any discussion in the literature of how the class fraction specifically affects the model-based priors.

[7]If this were the case, we might be tempted to address the class imbalance problem by setting the distribution of the classes in the training data according to our priors. But this could be problematic for a rare class. For example, if we wanted 1000 times as many stars as galaxies, but were limited to just 10 000 training vectors, our training data set would have to have only 10 galaxies, hardly a good basis with which to learn the characteristics of galaxies.

[8]In general we do not need a test set to calculate this. Only in the modified case do we need to know the true classes to calculate $P(x_n|\theta)$ (see section 5.1).

| | $G$ | star | physbin | quasar | galaxy |
|---|---|---|---|---|---|
| $P(C_j\|\theta^{nom})$ | 15 | 0.2508 | 0.2492 | 0.2499 | 0.2501 |
| $P(C_j\|\theta^{nom})$ | 18.5 | 0.2545 | 0.2414 | 0.2541 | 0.2500 |
| $P(C_j\|\theta^{nom})$ | 20 | 0.2598 | 0.2391 | 0.2537 | 0.2474 |
| $f_i^{nom}$ | all | 0.2500 | 0.2500 | 0.2500 | 0.2500 |

The bottom row gives, for comparison, the relative sizes of the class populations in the test data, i.e. the fraction of objects in each true class, $i$. This is the "nominal" case, i.e. the distribution on which the model was trained.[9] At least for this SVM model with equal class fractions, the model-based priors are approximately equal to the class fractions.

We will now see how we can modify the priors and thus apply DSC to a situation in which we expect a different class distribution from the one we trained it on.

# 5 Modifying the posterior probabilities to account for modified class fractions or priors

Often we train a classification algorithm on more or less equal proportions of each class, the idea being that this helps the algorithm to reliably identify the boundaries or fit the data/class distributions. Yet in the real world the true class fractions may be very different. For example, with Gaia we "know" (or rather, our prior is) that quasars are much rarer than stars (see section A). A DSC trained on equal proportions of classes may make too many false positive detections of quasars, resulting in a contamination higher than predicted (see section 3). The problem is that the model-based priors differ from our (desired) priors. We therefore need to modify the DSC output probabilities to account for modified class fractions, or rather for the modified priors, to give us the "modified model", $\theta^{mod}$. From inspection of equation 2 we see that this could be achieved by modifying the priors directly, i.e. by dividing the posterior by the original (nominal) prior and then multiplying it by the modified prior. The nominal priors could be calculated as outlined in section 4. However, this cannot be used to calculate the modified priors because we don't yet know the posteriors from the modified model. We therefore approximate the priors using the class fractions.[10]

Let $P^{nom}(C_j|x,\theta^{nom})$ be the output (posterior) probability from DSC for object $x$ for output class $j$ in the case that the nominal class fractions are $f_i^{nom}$ (the fraction of sources in the training set

[9] These calculations are made using a test data set with the same class distribution as the training data set. We could use the training data themselves, but it's better to use a test set to avoid biases from possible overfitting.

[10] On the basis of the discussion in section 4 we would not expect this always to be a good approximation. But in some cases it is (as we see for the nominal case, section 4.3) and we can at least use equation 3 to check the assumption.

of true class $i$). The modified probability when the class fractions would be $f_i^{mod}$ is

$$P^{mod}(C_j|x_n,\theta^{mod}) = a_n P^{nom}(C_j|x_n,\theta^{nom}) \frac{f_{i=j}^{mod}}{f_{i=j}^{nom}} \tag{4}$$

where $a_n$ is a normalization factor which is calculated to ensure that $\sum_j P^{mod}(C_j|x_n,\theta^{mod}) = 1$ for each object $n$. Varying the ratio $\frac{f_{i=j}^{mod}}{f_{i=j}^{nom}}$ has the expected impact on the posteriors, although the need to renormalize is important and its impact may be less obvious. It may be tempting to think of these modified probabilites as those we would have achieved *if* we had trained the model on the modified class fractions. But this is generally *not* true, because what we are doing in equation 4 is changing the priors and this is not equivalent to changing the training data, as discussed in section 4.2.

The equations for the completeness and contamination (eqn. 1) of a selected sample depend on the number of objects correctly and incorrectly classified. They are calculated from the test data set, which will typically have the same class fraction as the training data set, the nominal class fractions, $f_{i=j}^{nom}$. However, we expct to apply the modified model to a data set with different class fractions (this was the whole point of modifying it), so the number of objects of each true class would be different. To predict the the completeness and contamination in that case we must modify the equations to be

$$\begin{aligned} completeness_j &= \frac{\left(\frac{f_{i=j}^{mod}}{f_{i=j}^{nom}}\right) n_{i=j,j}}{\left(\frac{f_{i=j}^{mod}}{f_{i=j}^{nom}}\right) N_i} = \frac{n_{i=j,j}}{N_i} \\ contamination_j &= \frac{\sum_{i\neq j}\left(\frac{f_i^{mod}}{f_i^{nom}}\right) n_{i,j}}{\sum_i \left(\frac{f_i^{mod}}{f_i^{nom}}\right) n_{i,j}} \end{aligned} \tag{5}$$

The quantities $n$ and $N$ still refer to the nominal data set. Note that the class fractions are normalized (i.e. $\sum_i f_i^{nom} = \sum_i f_i^{mod} = 1$). Note also that the *equation* for the completeness remains unchanged from the nominal case, yet the *value* of the completeness will generally change because the posterior probabilities have changed and so will the sample selected by applying a threshold. Of course, to calculate the completeness and contamination for an actual test set which really does have the class fractions we want, we would use equation 1.

## 5.1 Calculating the model-based priors from the modified model

In the previous section we used the modified class fractions, $f_i^{mod}$, as a proxy for the modified priors in order to calculate the modified posteriors. But now that we have these posteriors, we can calculate the model-based priors for the modified model using equation 3 and compare to $f_i^{mod}$. With the nominal model $P(x_n|\theta^{nom})$ in equation 3 was just $1/N$: each object in the

training set only appeared once. In the modified model we are effectively changing the relative number of objects of each class, so these probabilities must be modified consistently with this to give

$$P(x_n|\theta^{mod}) = \left(\frac{f_i^{mod}}{f_i^{nom}}\right)\frac{1}{N} \tag{6}$$

where $x_n$, being an object in the (labelled) test set, has known true class, $i$. If a class is ten times rarer in the modified population, $P(x_n|\theta^{mod})$ is reduced by a factor of (approximately) ten: the prior for that object is (approximately) ten times smaller.[11]

Applying this to the three modified data sets used in section 6 (without astrometry) we obtain the following

| | $G$ | star | physbin | quasar | galaxy |
|---|---|---|---|---|---|
| $P(C_j\|\theta^{mod})$ | 15 | 0.3350 | 0.3320 | 0.0002579 | 0.3327 |
| $P(C_j\|\theta^{mod})$ | 18.5 | 0.3459 | 0.3207 | 0.0002837 | 0.3332 |
| $P(C_j\|\theta^{mod})$ | 20 | 0.3516 | 0.3187 | 0.0002831 | 0.3295 |
| $f_i^{mod}$ | all | 0.3332 | 0.3332 | 0.0003332 | 0.3333 |

The agreement with the modified class fractions (bottom line) is good and implies that the use of the class fractions in equation 4 to estimate the modified posteriors is valid.

# 6 Results

Separate DSC models are trained and tested on objects with a common magnitude. Four sets of results are presented here: G=15 with astrometry; G=18.5 with and without astrometry; G=20 with astrometry. At G=15 we don't expect many quasars; this is included because the noise is very low, i.e. a limiting/ideal case. Results are presented for the "nominal" and "modified" cases. In the nominal cases, the relative class fractions are equal, i.e. relative fractions of $(1,1,1,1)$. In the modified case, the DSC output probabilities are modified (in the way described in section 5) to accommodate different expected fractions (priors). Specifically, quasars are made 1000 times rarer, so that the modified class fraction vector is $(1,1,0.001,1)$. This is the order of the star/quasar ratio we expect for Gaia at G=20 (but see section A). Of course, the binary and galaxy ratio would also need to be modified for a complete realistic classifier, but for this first investigation we vary just a single class so as to ease interpretation.

Results are presented in three ways. The first is a contingency table, or confusion matrix. This shows the correct classifications (on-diagonal values) and misclassification values (off-diagonal

[11] Putting equations 6 and 4 into equation 3 we see that the factor $\frac{f_i^{mod}}{f_i^{nom}}$ appears squared. But this is correct, as the model-based priors depend on the product of two factors, both of which depend on the training data class distribution.

values) as percentages. An object is assigned to that class for which the corresponding DSC output is largest. This is only one way of assigning classes and thus deriving classification accuracies and is not what we want when trying to get pure samples or when modifying the priors. Instead we want to assign thresholds (see section 3). We therefore show various histograms of the DSC output probabilities which tell us the confidence with which classes are assigned. For each of the four data sets and nominal/modified cases we show histograms of: $P(\texttt{class}|\texttt{CLASS})$, showing how confident the true positives are (e.g. Fig. 6); $P(\texttt{class}|\texttt{QUASAR})$, showing the probabilities assigned to true quasars (e.g. Fig. 7); $P(\texttt{quasar}|\texttt{CLASS})$, showing the quasar probabilities assigned to objects of each true class (e.g. Fig. 8); Using the posteriors to build samples, we then look at how the sample completeness and contamination varies with the threshold used. In all cases completeness will decrease monotonically from 1 at $P_t = 0$ (all objects included in sample) to 0 to $P_t = 1$ (no objects in sample). The maximum contamination will occur at $P_t = 0$ with a value depending on the class fractions. For equal numbers of objects in $N$ classes it will be $1/N$. Contamination will similarly drop to zero at $P_t = 1$, but a little thought will convince the reader that it need not decrease monotonically with increasing $P_t$.

All histograms are density estimates with total area unity. They were created using the `truehist` function in the `MASS` package in **R**.

## 6.1 G=15

Table 1: Contingency table. Each row corresponds to a true class and thus sums to 100%. Nominal priors, G=15

| | galaxy | physbinary | quasar | star |
|---|---|---|---|---|
| GALAXY | 99.95 | 0.00 | 0.00 | 0.05 |
| PHYSBINARY | 0.00 | 96.85 | 0.00 | 3.15 |
| QUASAR | 1.05 | 0.00 | 97.60 | 1.35 |
| STAR | 0.00 | 6.90 | 0.35 | 92.75 |

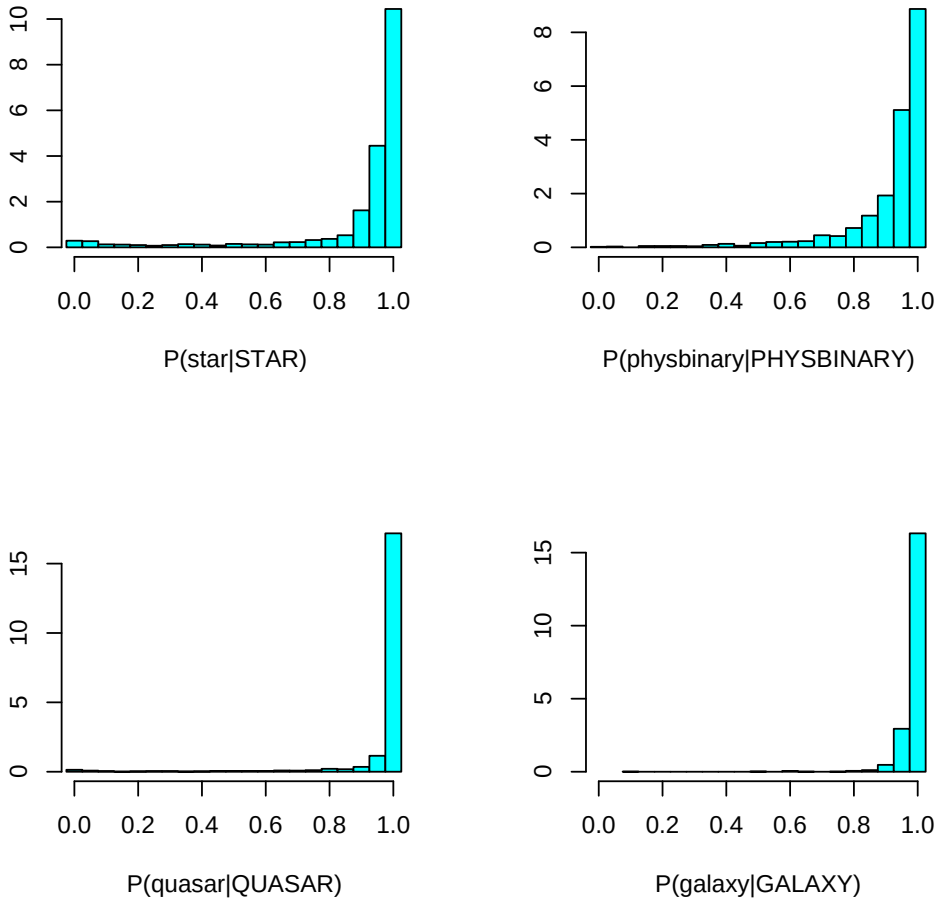


Figure 6: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Nominal priors, G=15

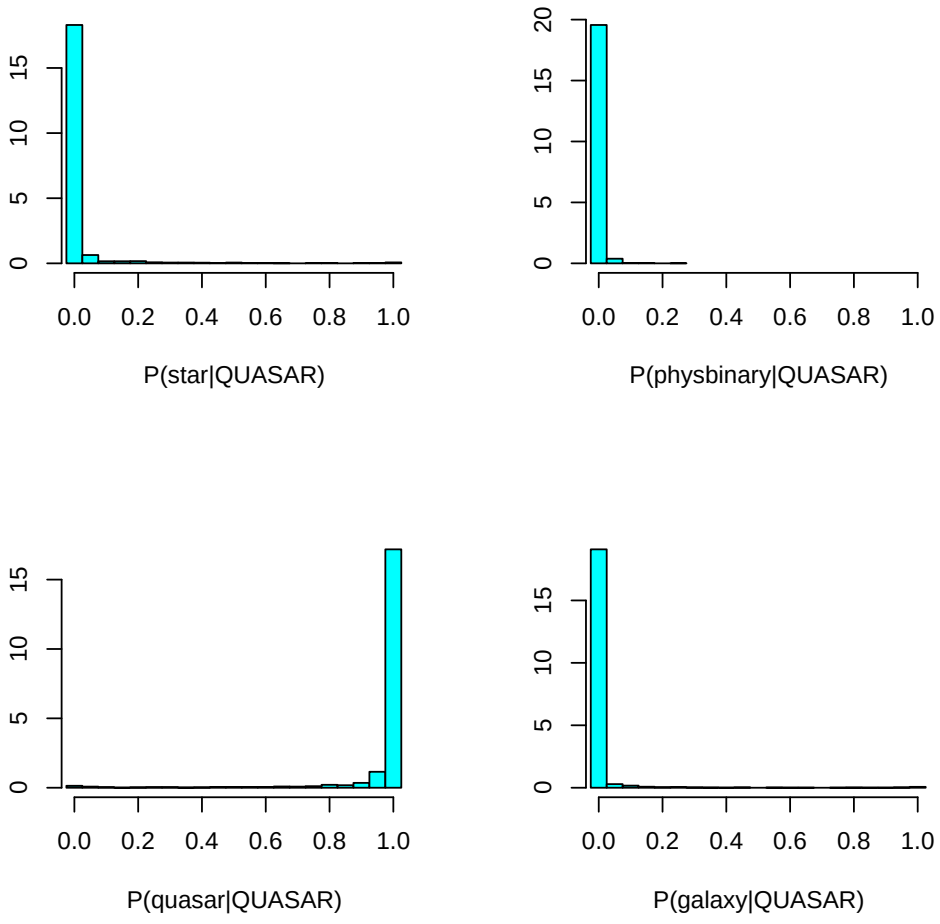


Figure 7: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Nominal priors, G=15

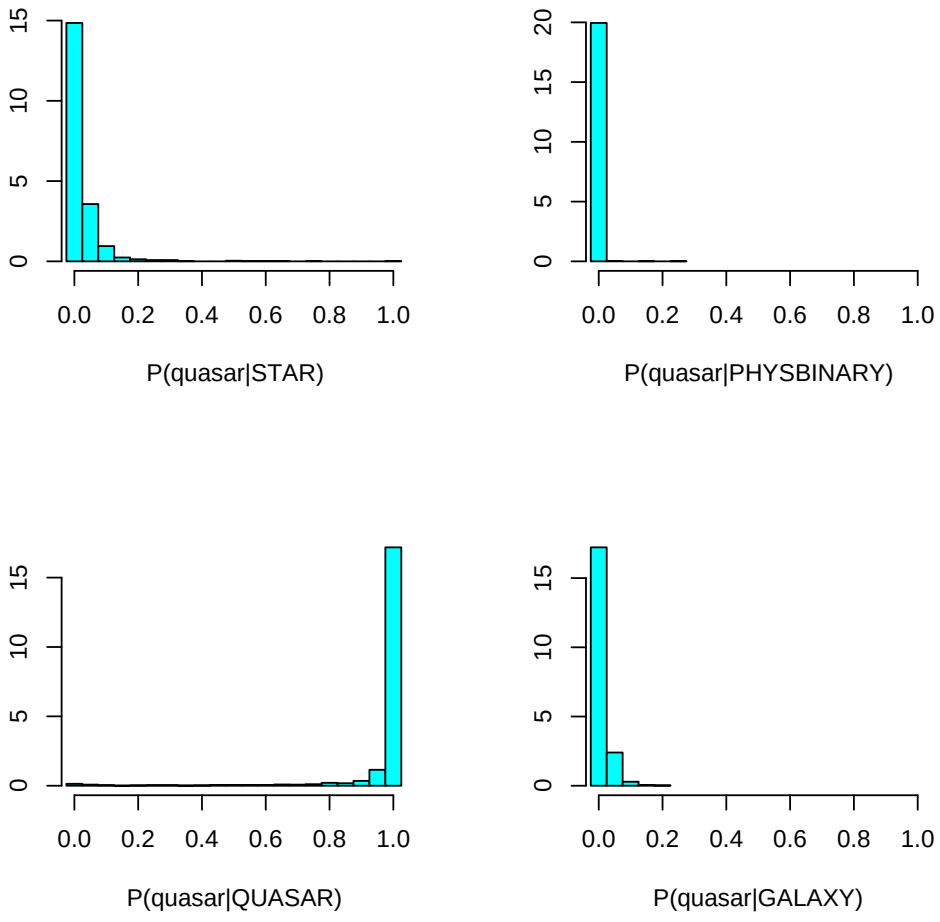


Figure 8: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Nominal priors, G=15

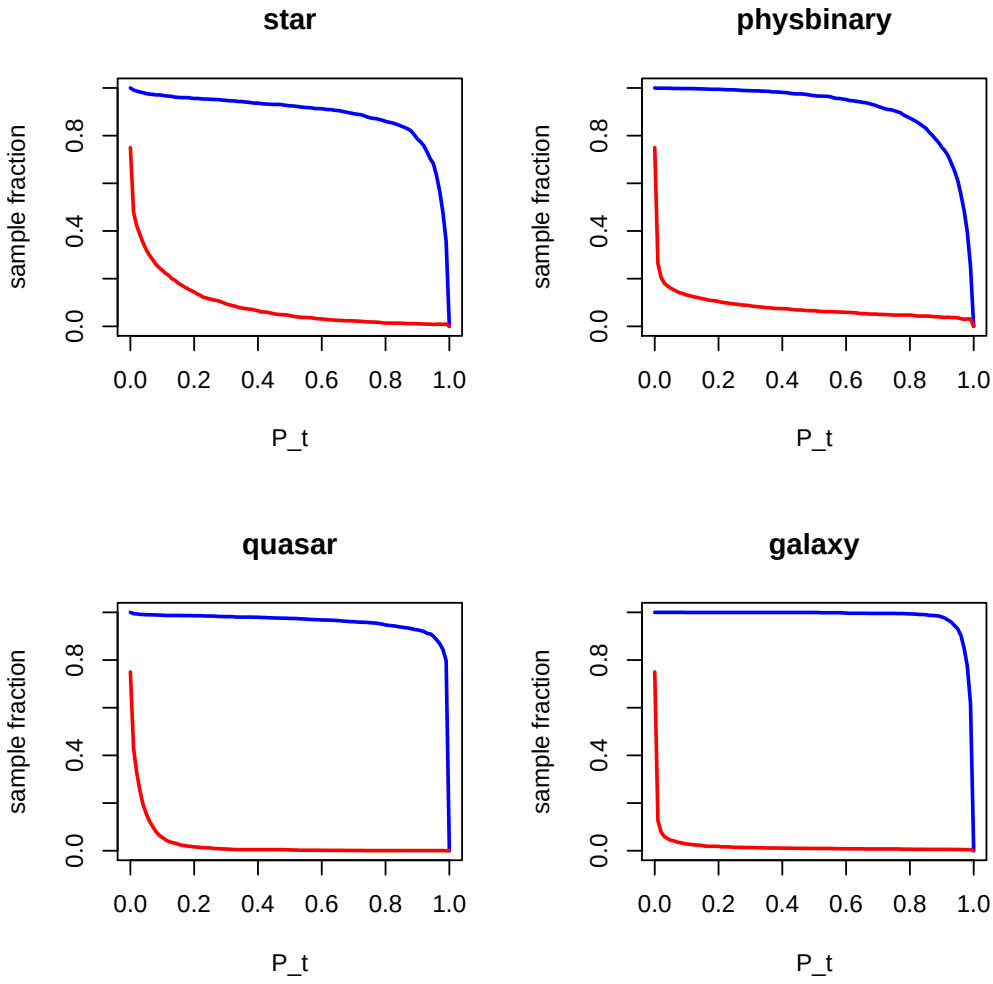


Figure 9: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Nominal priors, G=15

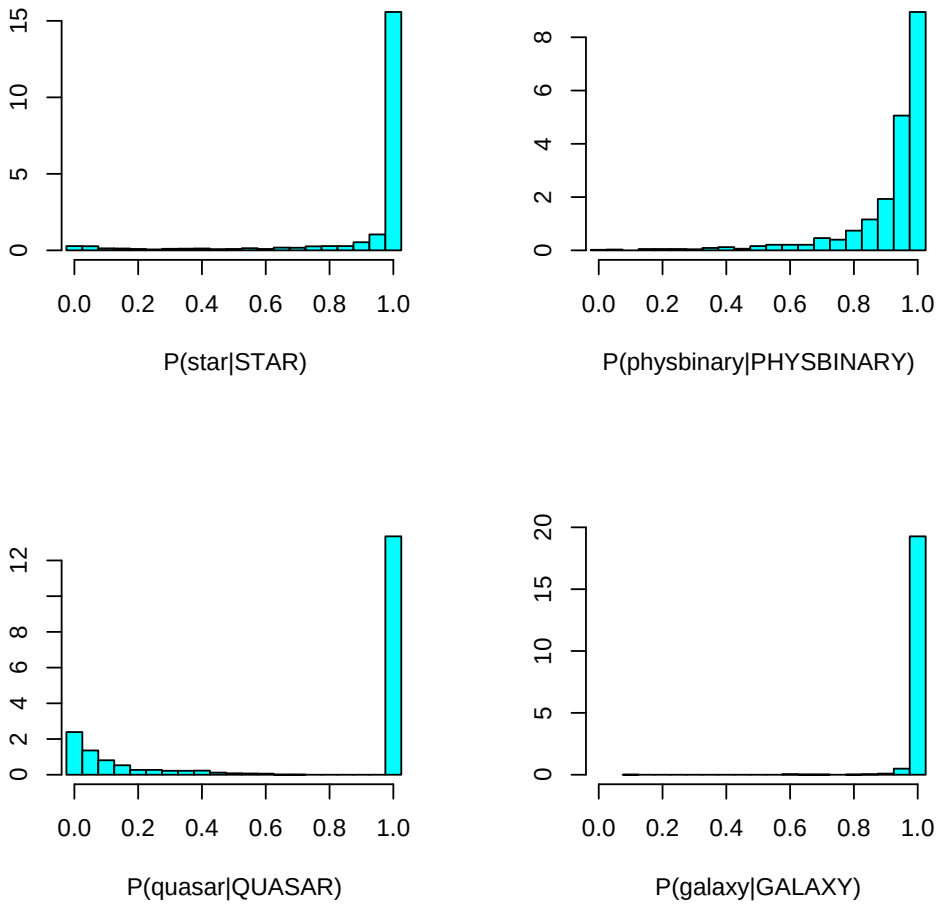


Figure 10: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Modified priors, G=15

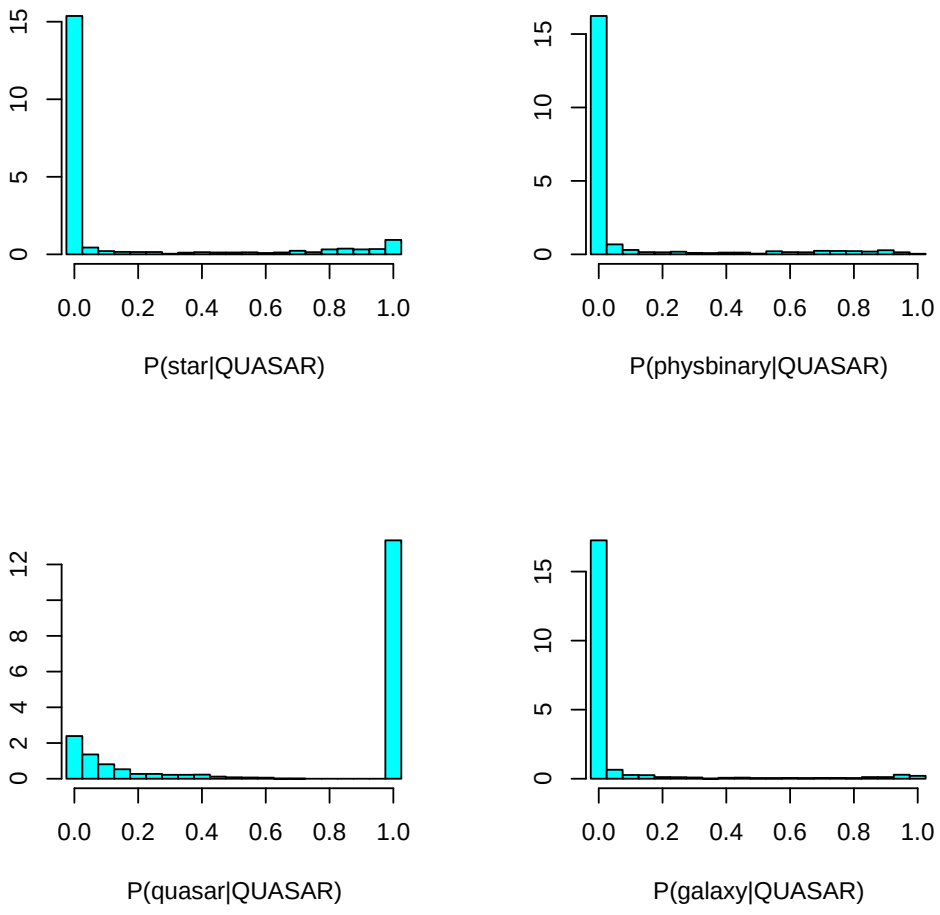


Figure 11: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Modified priors, G=15

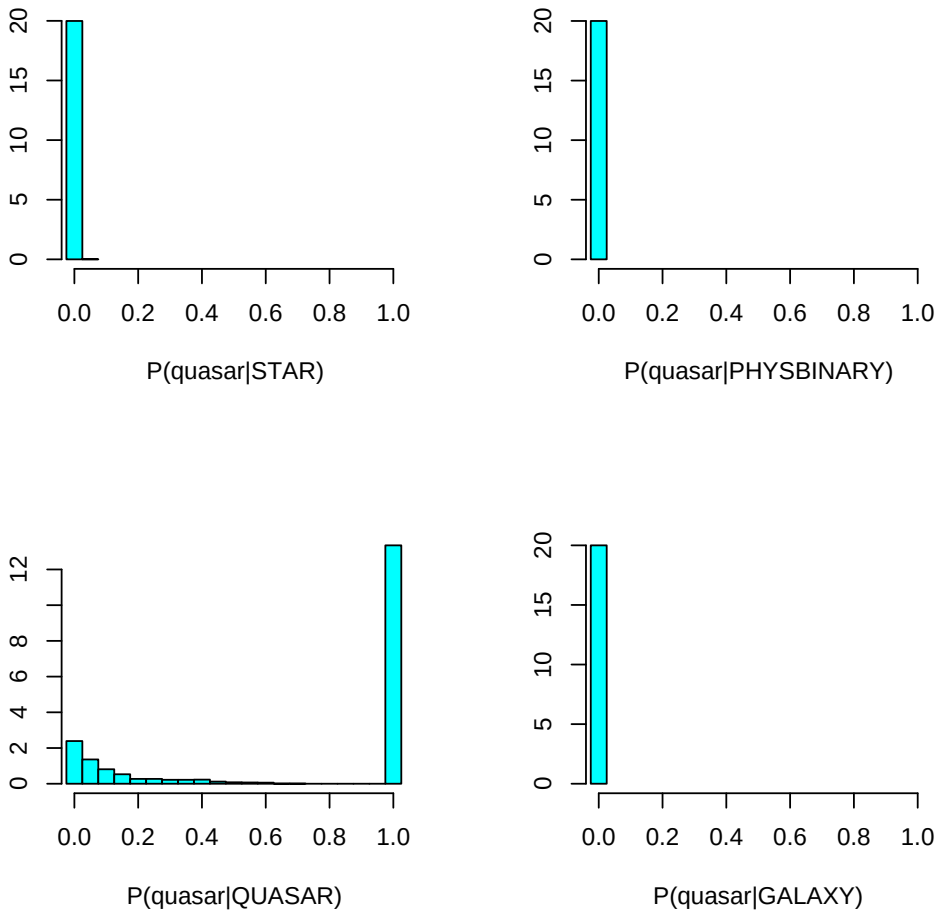


Figure 12: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Modified priors, G=15

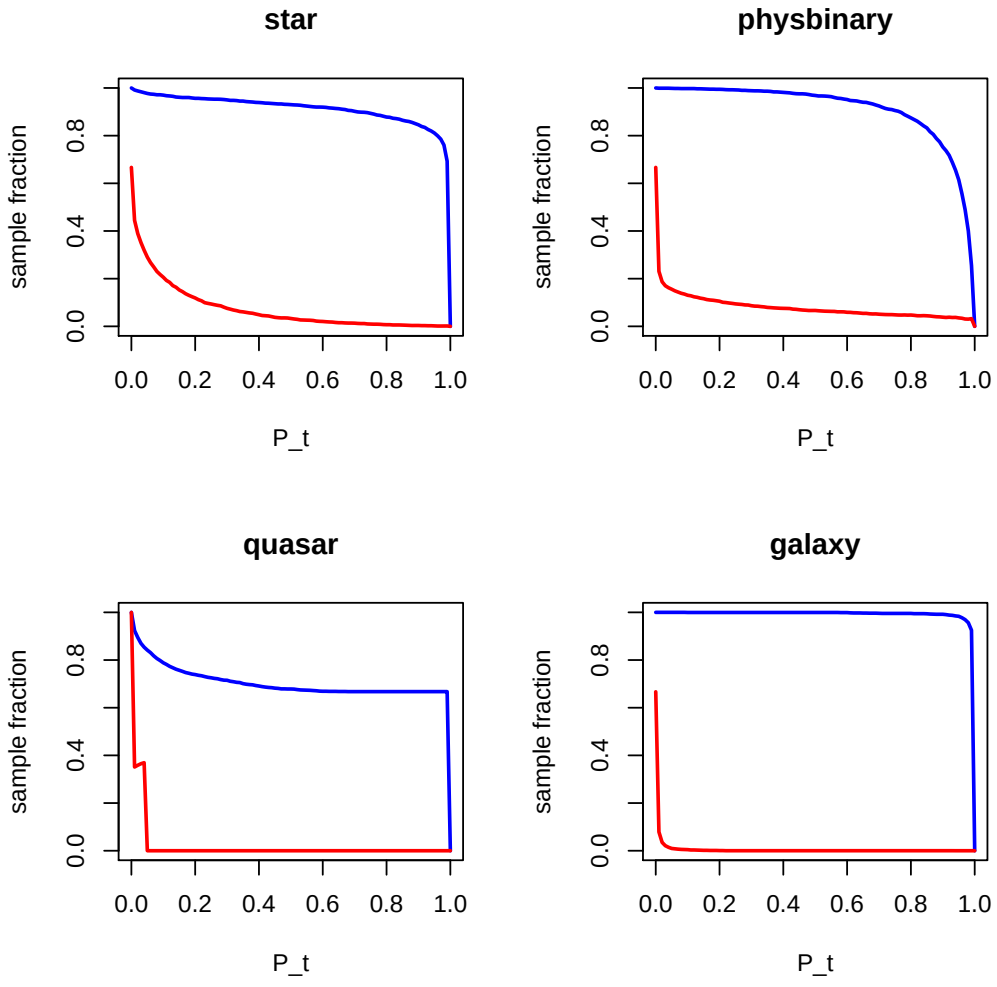


Figure 13: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Modified priors, G=15

## 6.2 G=18.5

Table 2: Contingency table. Each row corresponds to a true class and thus sums to 100%. Nominal priors, G=18.5

| | galaxy | physbinary | quasar | star |
|---|---|---|---|---|
| GALAXY | 96.35 | 0.05 | 1.30 | 2.30 |
| PHYSBINARY | 0.25 | 86.85 | 0.10 | 12.80 |
| QUASAR | 3.80 | 0.65 | 88.25 | 7.30 |
| STAR | 1.65 | 13.00 | 6.70 | 78.65 |

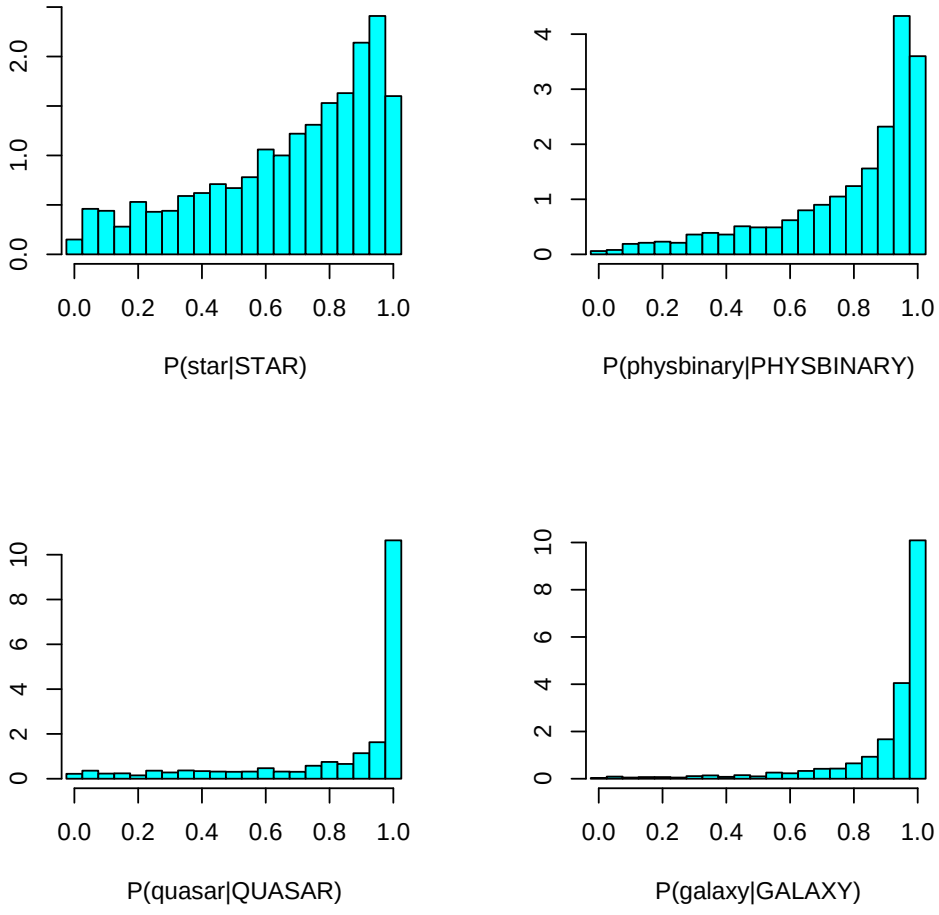


Figure 14: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Nominal priors, G=18.5

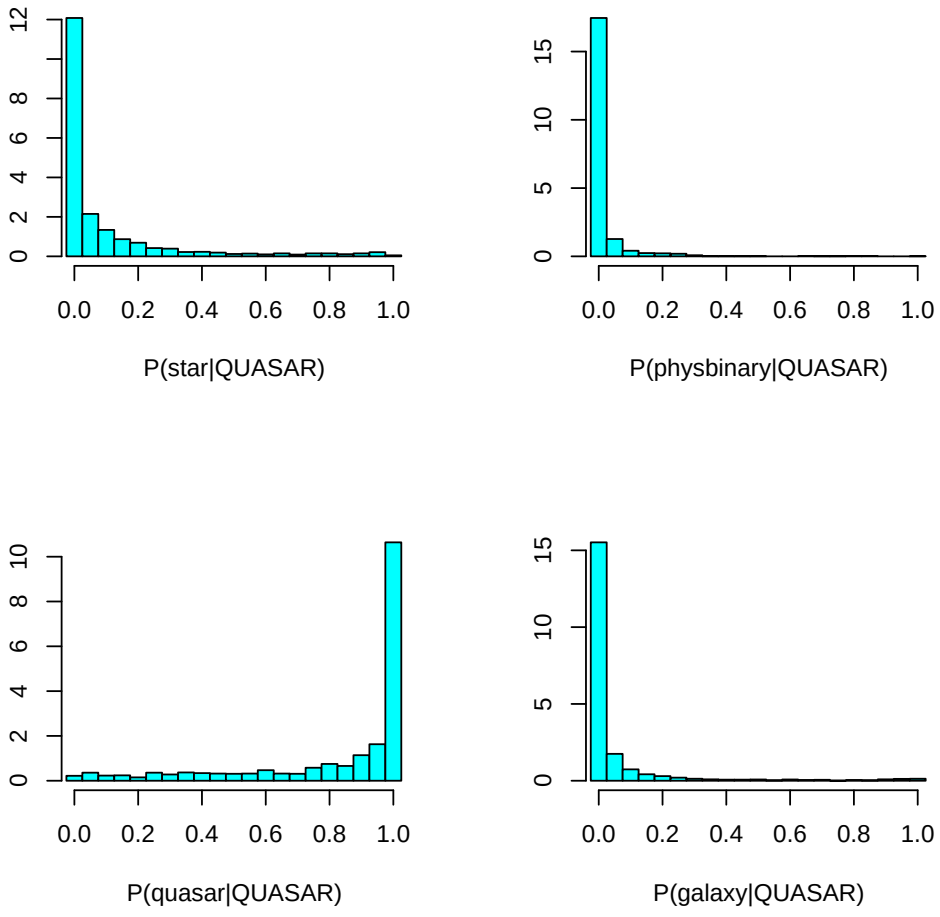


Figure 15: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Nominal priors, G=18.5

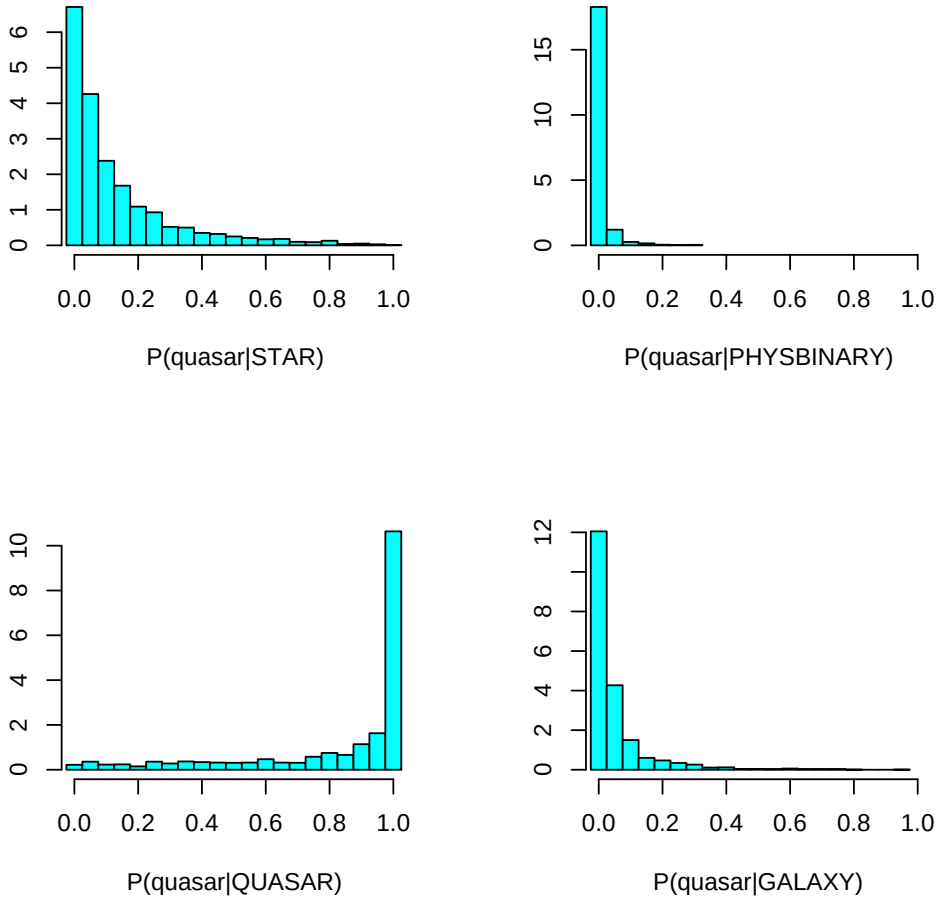


Figure 16: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Nominal priors, G=18.5

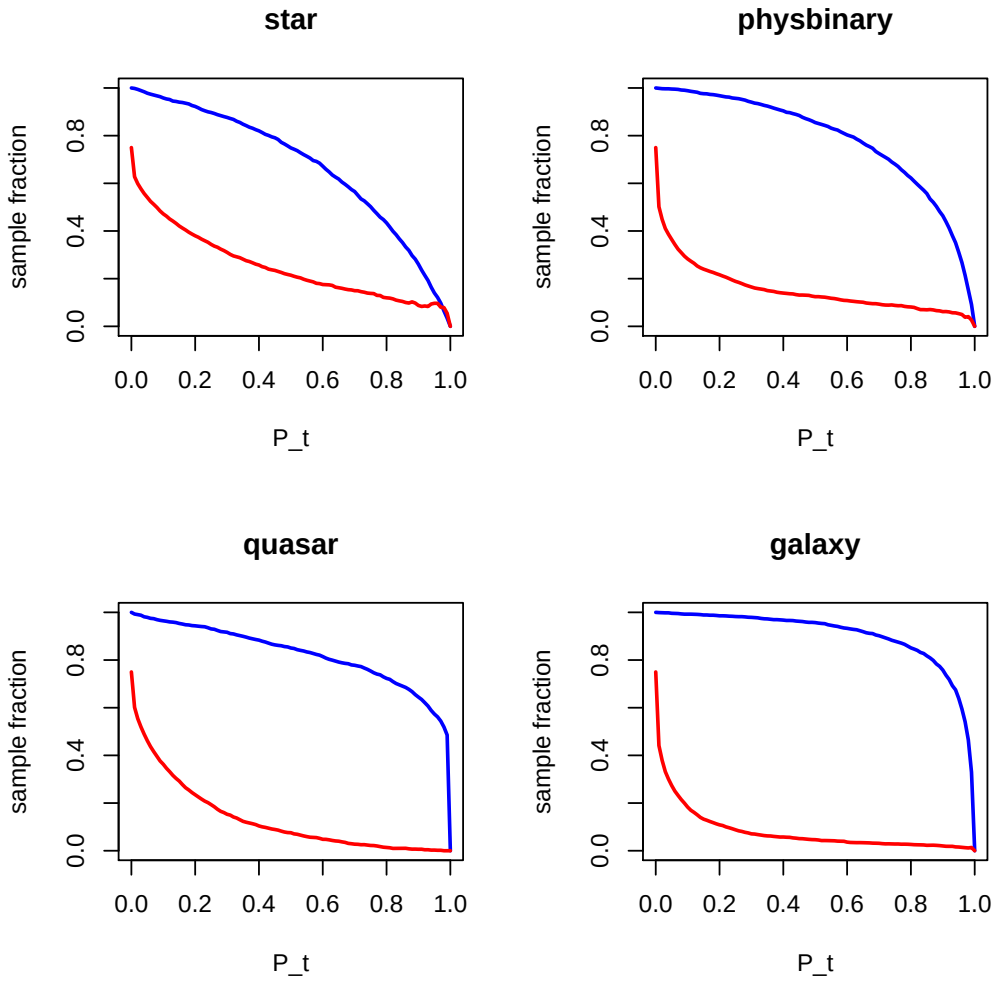


Figure 17: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Nominal priors, G=18.5

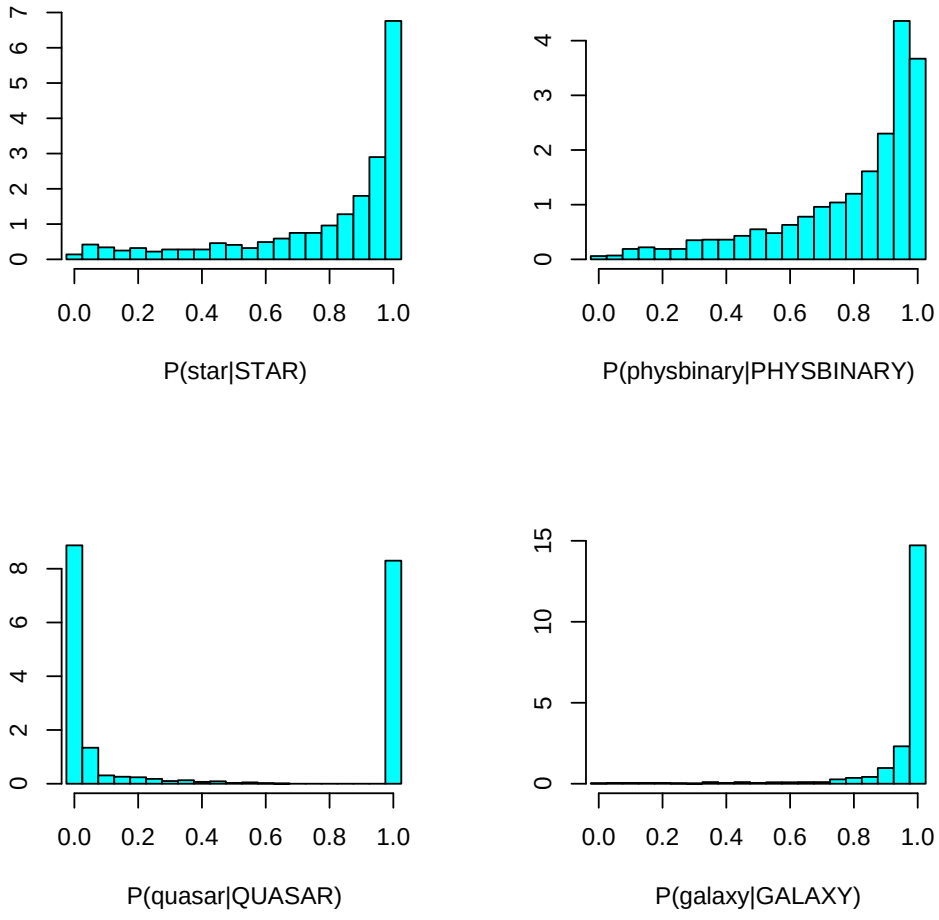


Figure 18: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Modified priors, G=18.5

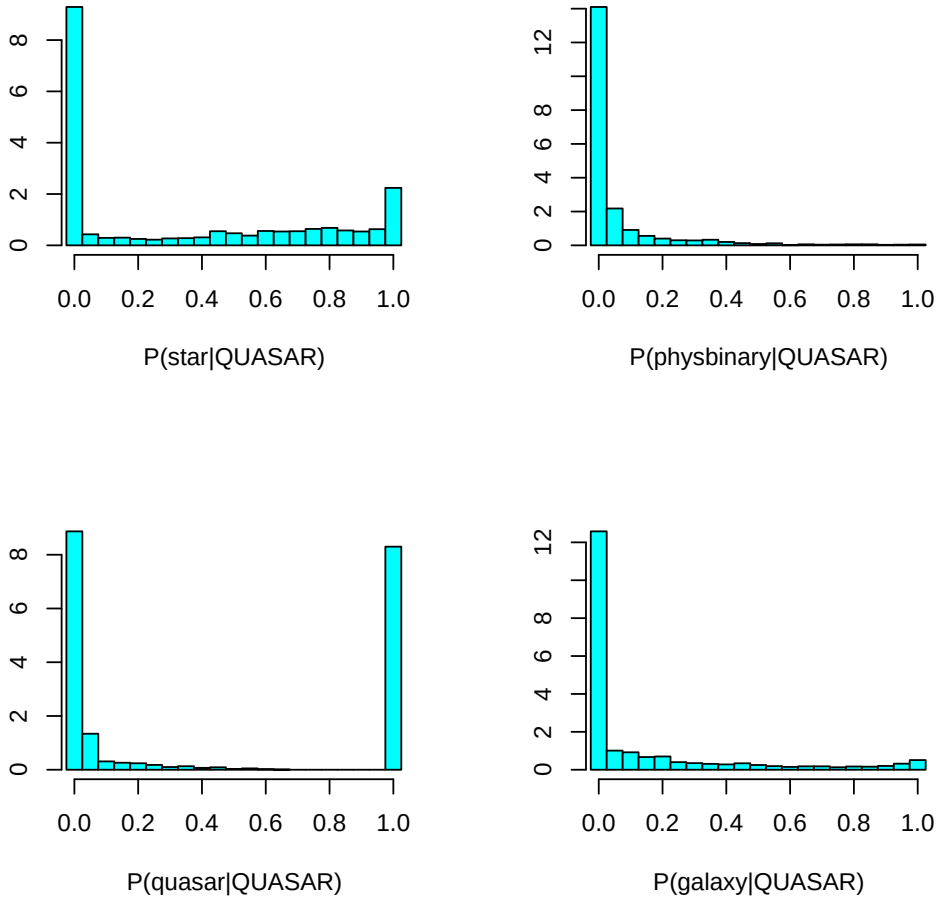


Figure 19: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Modified priors, G=18.5

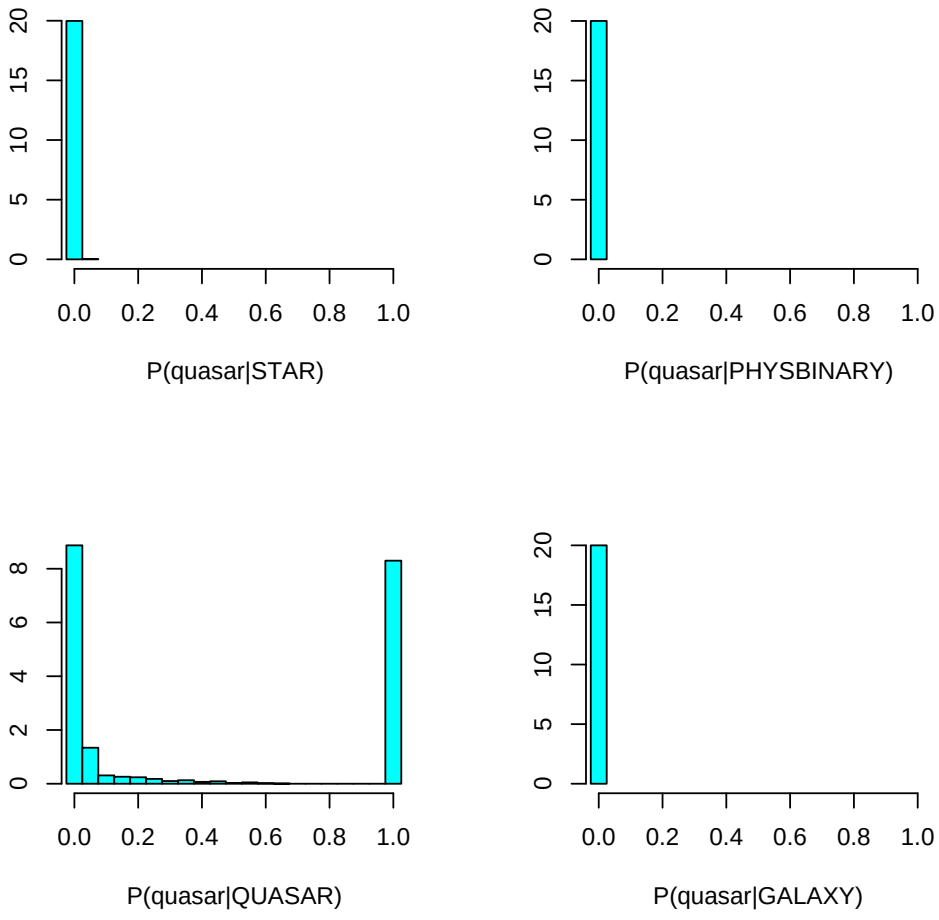


Figure 20: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Modified priors, G=18.5

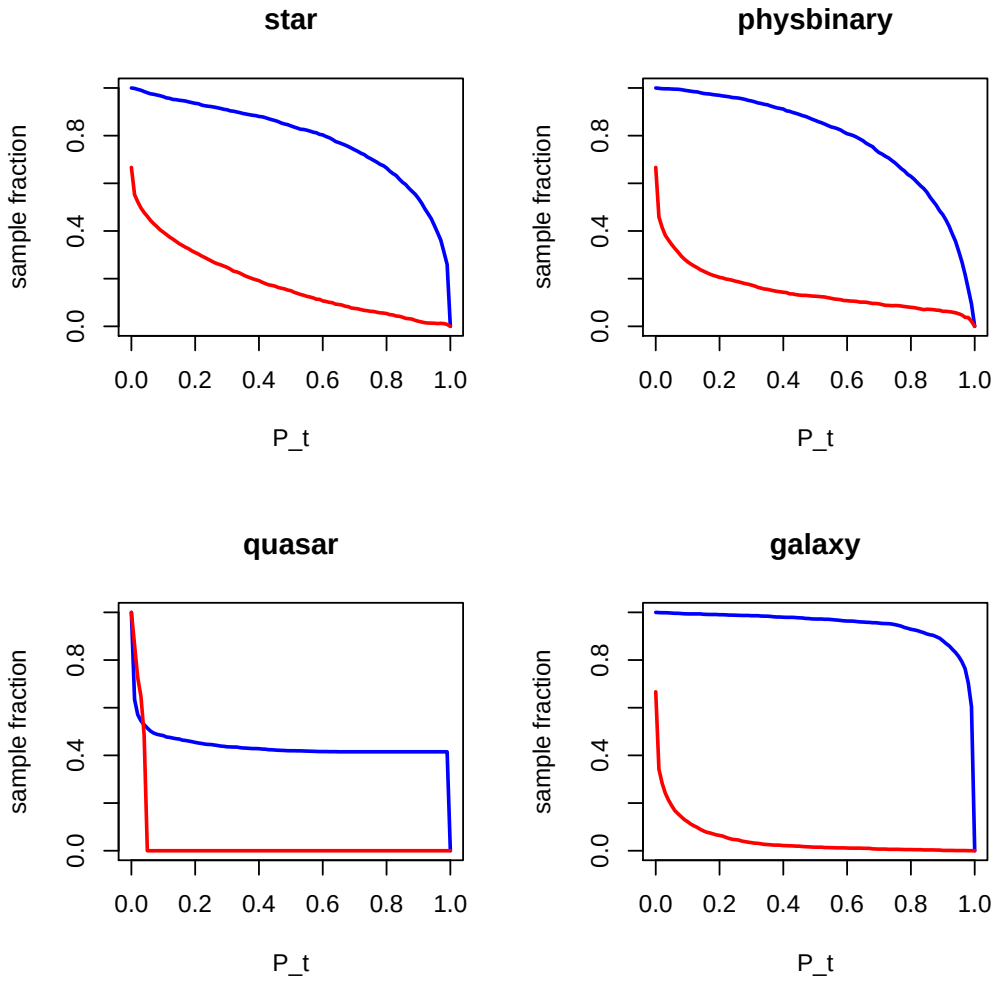


Figure 21: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Modified priors, G=18.5

## 6.3 G=18.5 without astrometry

Table 3: Contingency table. Each row corresponds to a true class and thus sums to 100%. Nominal priors, G=18.5 without astrometry

| | galaxy | physbinary | quasar | star |
|---|---|---|---|---|
| GALAXY | 96.10 | 0.20 | 1.35 | 2.35 |
| PHYSBINARY | 0.45 | 83.75 | 0.25 | 15.55 |
| QUASAR | 3.95 | 1.80 | 87.10 | 7.15 |
| STAR | 1.75 | 19.40 | 6.10 | 72.75 |

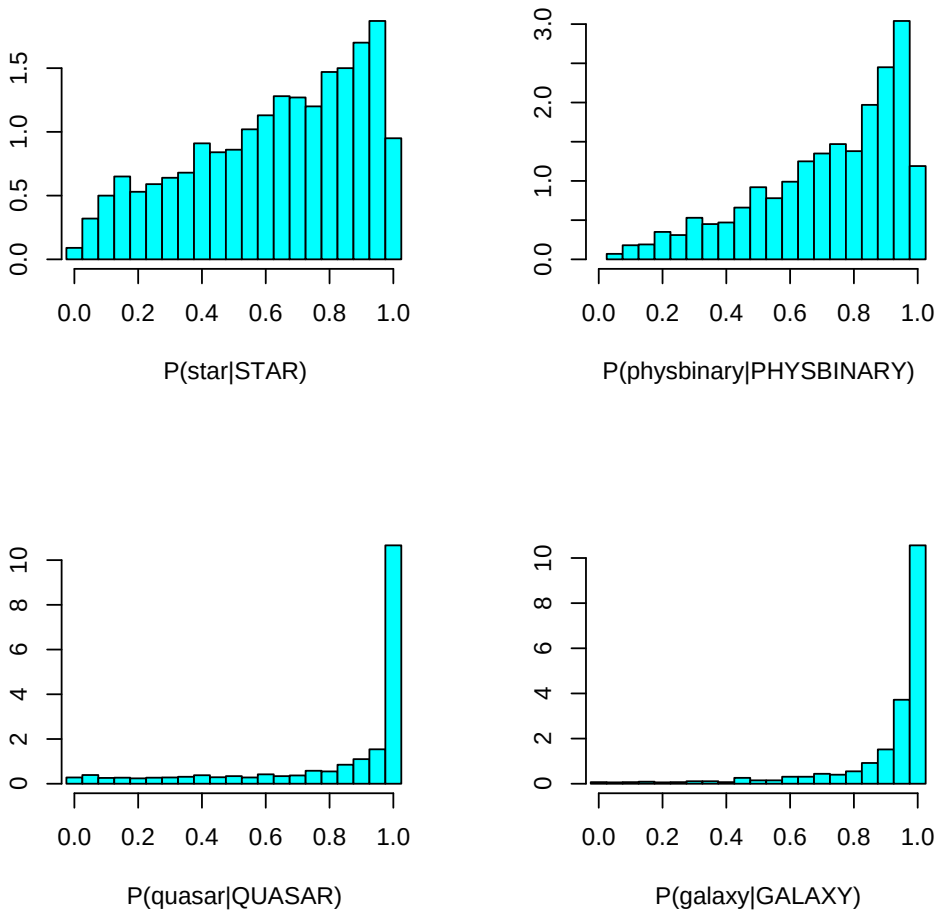


Figure 22: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Nominal priors, G=18.5, without astrometry

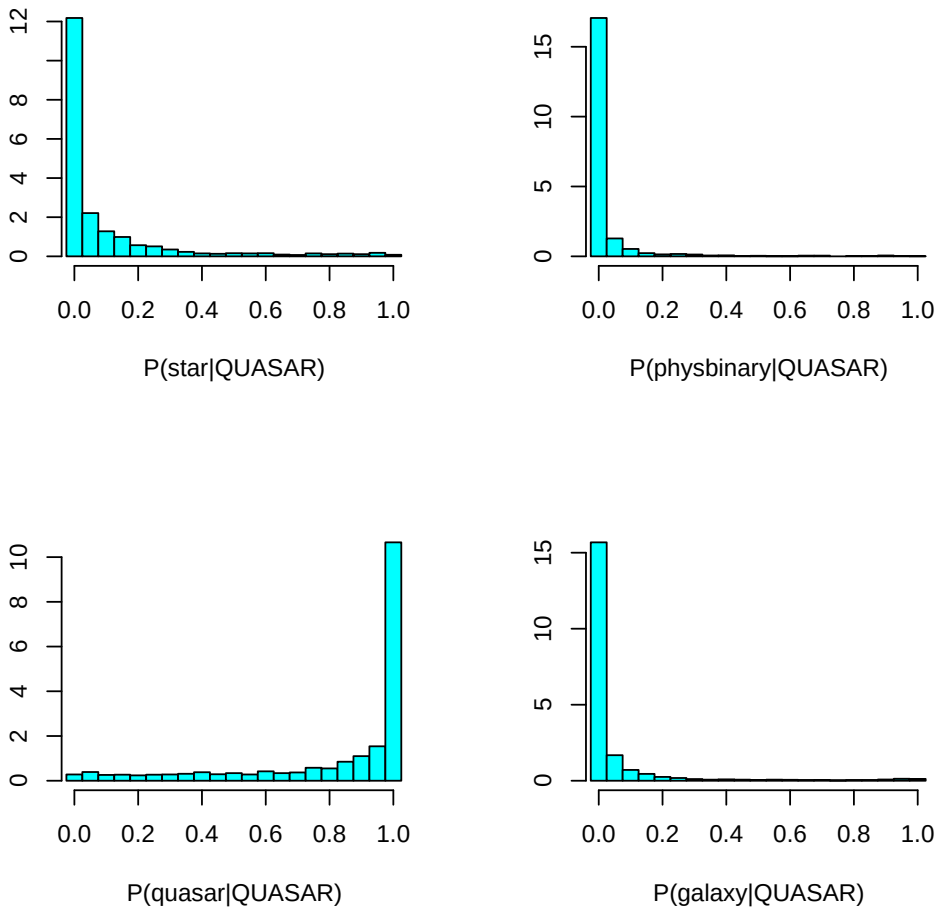


Figure 23: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Nominal priors, G=18.5, without astrometry

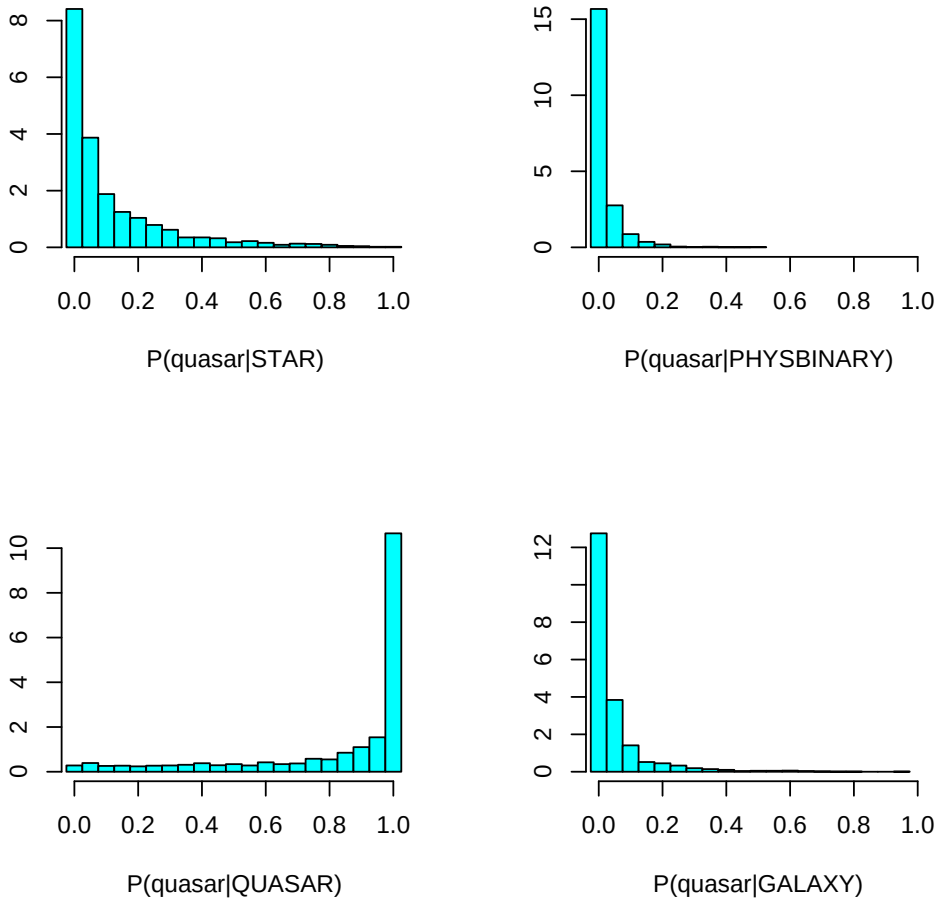


Figure 24: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Nominal priors, G=18.5, without astrometry

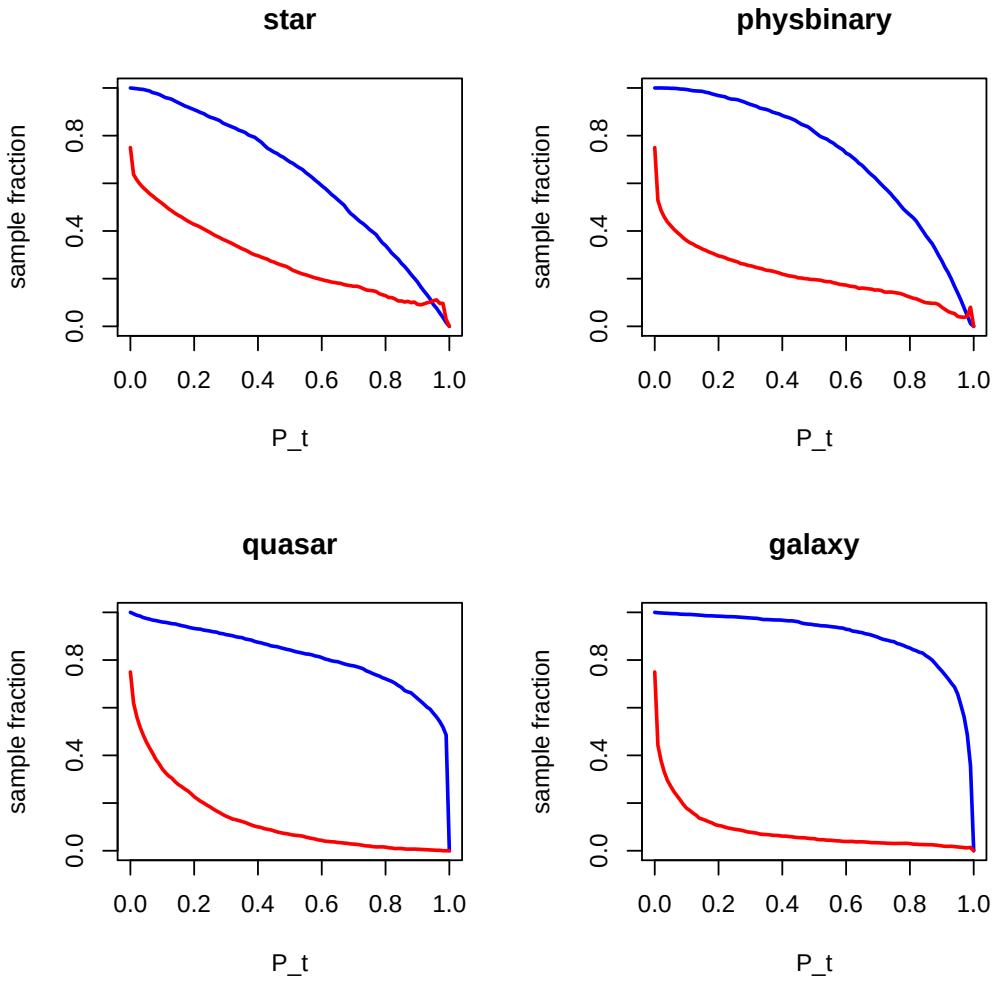


Figure 25: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Nominal priors, G=18.5, without astrometry

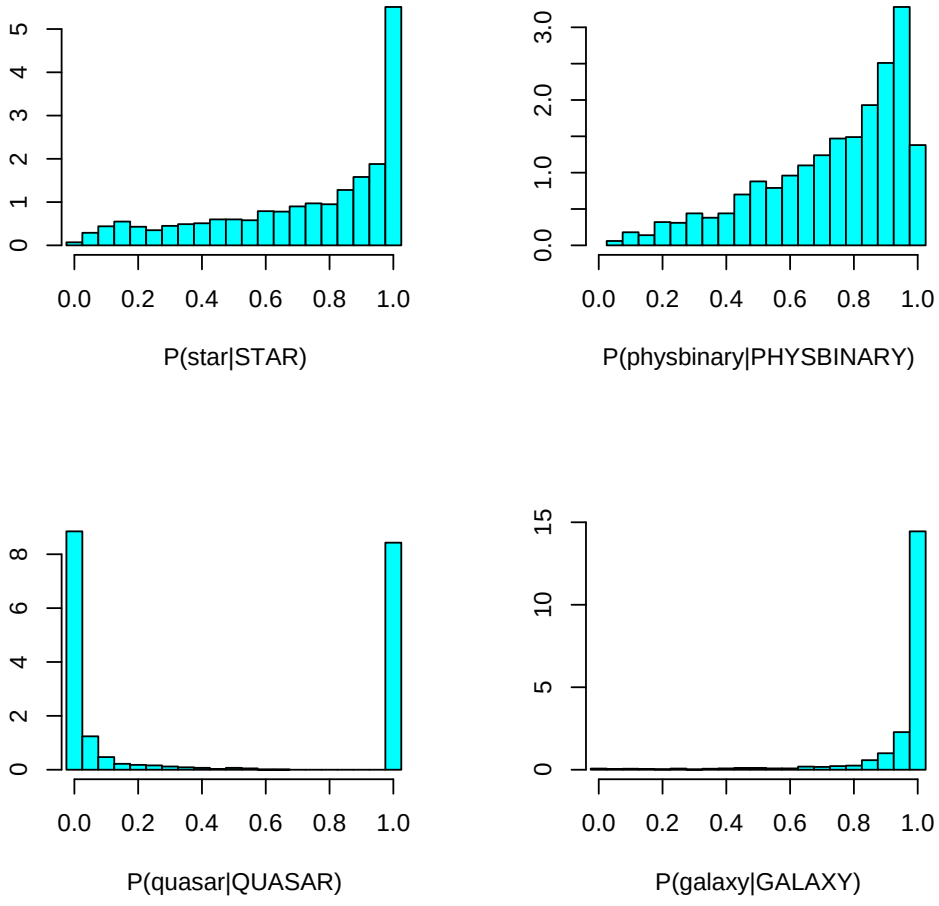


Figure 26: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Modified priors, G=18.5, without astrometry

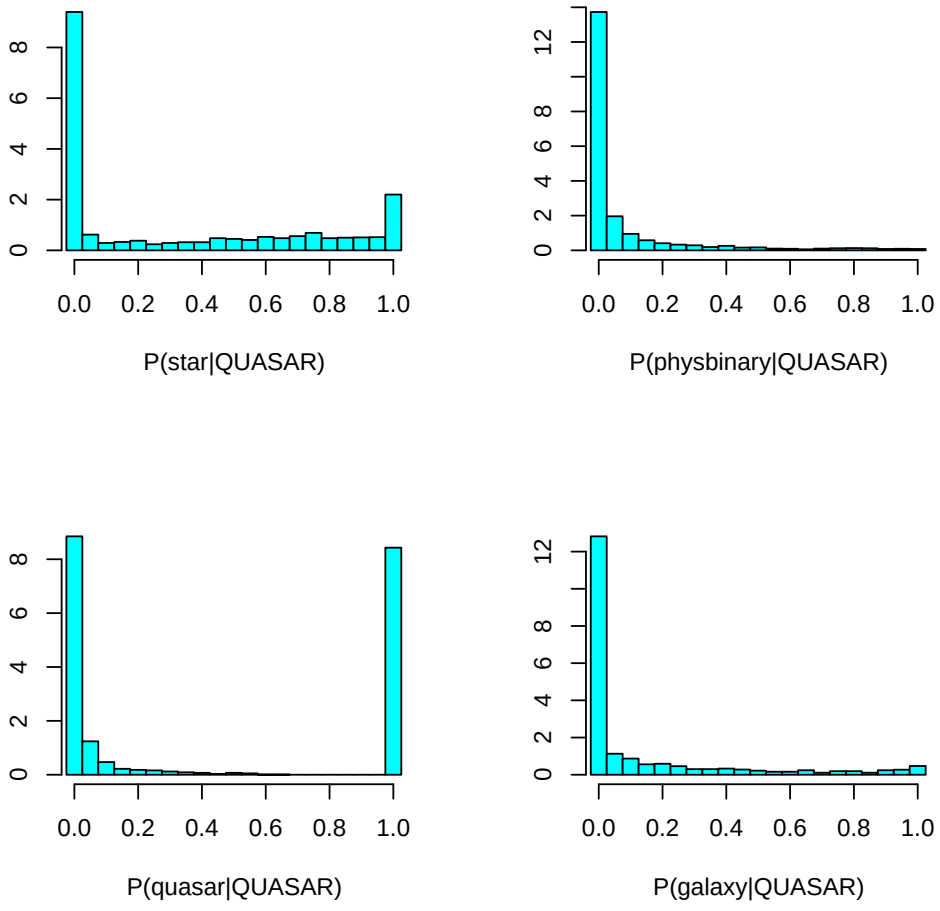


Figure 27: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Modified priors, G=18.5, without astrometry

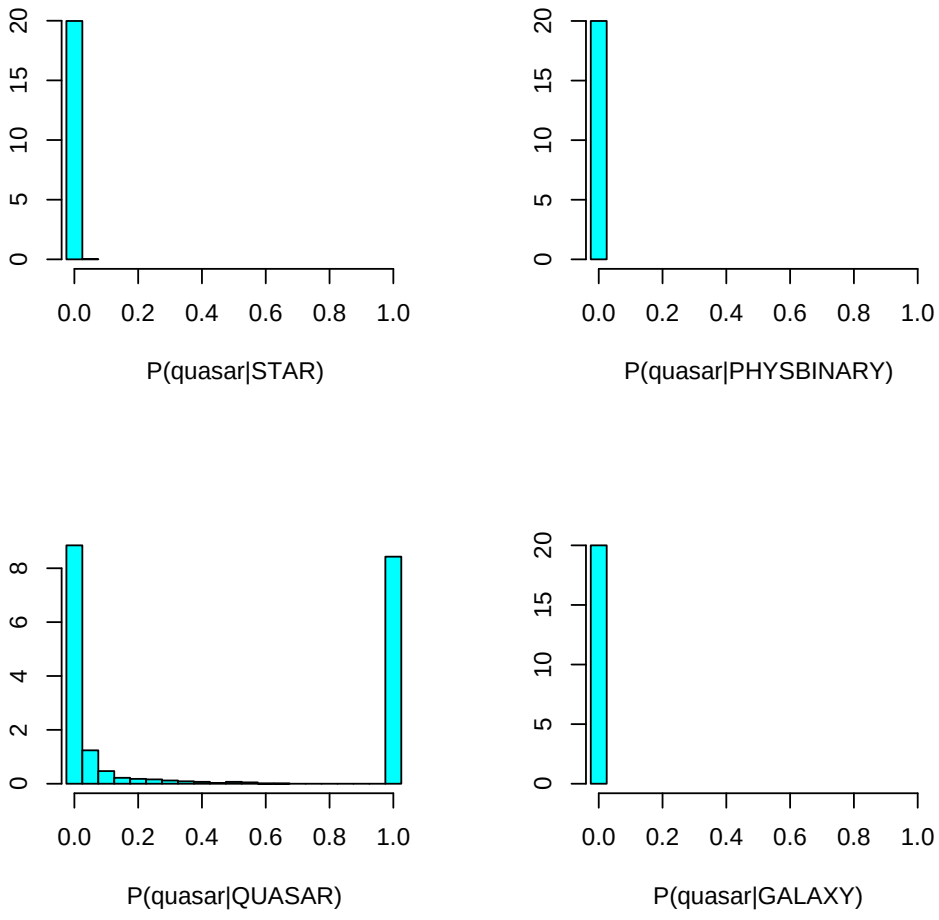


Figure 28: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Modified priors, G=18.5, without astrometry

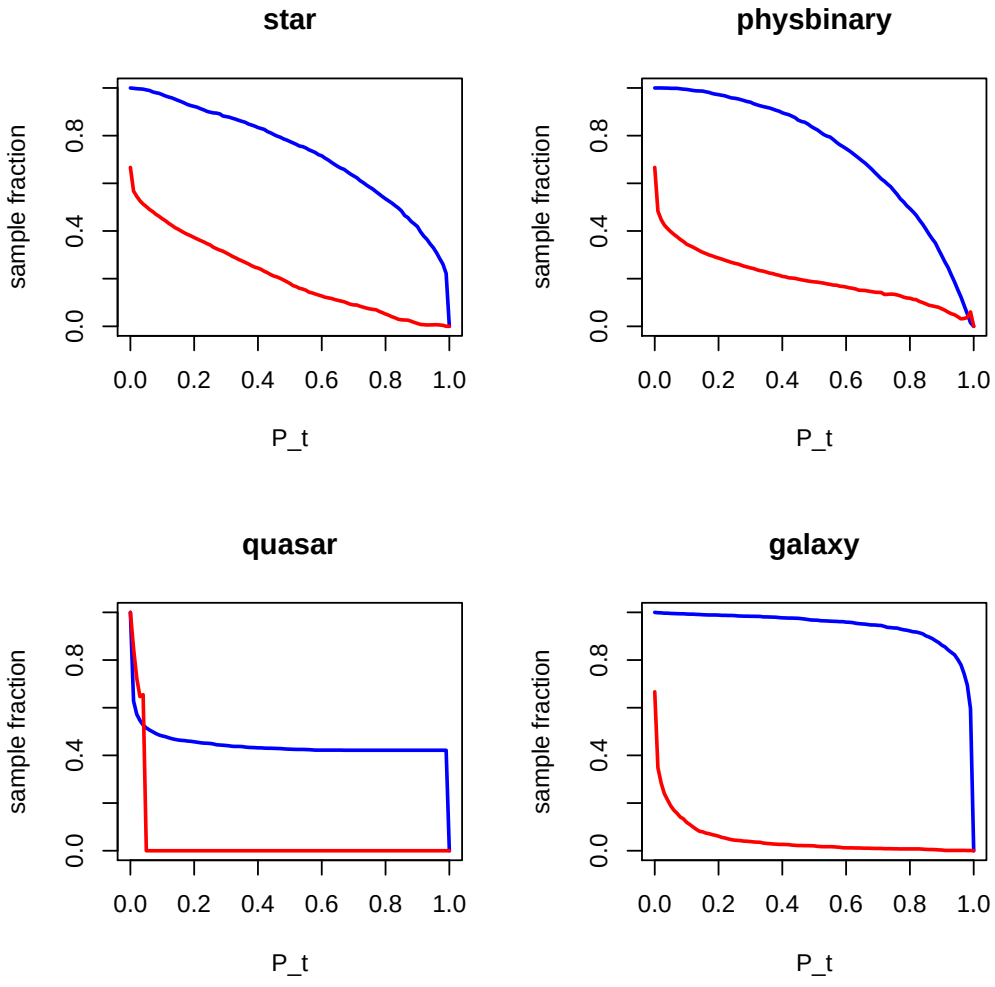


Figure 29: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Modified priors, G=18.5, without astrometry

## 6.4 G=20

Table 4: Contingency table. Each row corresponds to a true class and thus sums to 100%. Nominal priors, G=20

| | galaxy | physbinary | quasar | star |
|---|---|---|---|---|
| GALAXY | 88.30 | 1.70 | 2.85 | 7.15 |
| PHYSBINARY | 3.60 | 77.15 | 0.75 | 18.50 |
| QUASAR | 10.00 | 2.80 | 75.05 | 12.15 |
| STAR | 10.65 | 14.50 | 9.25 | 65.60 |

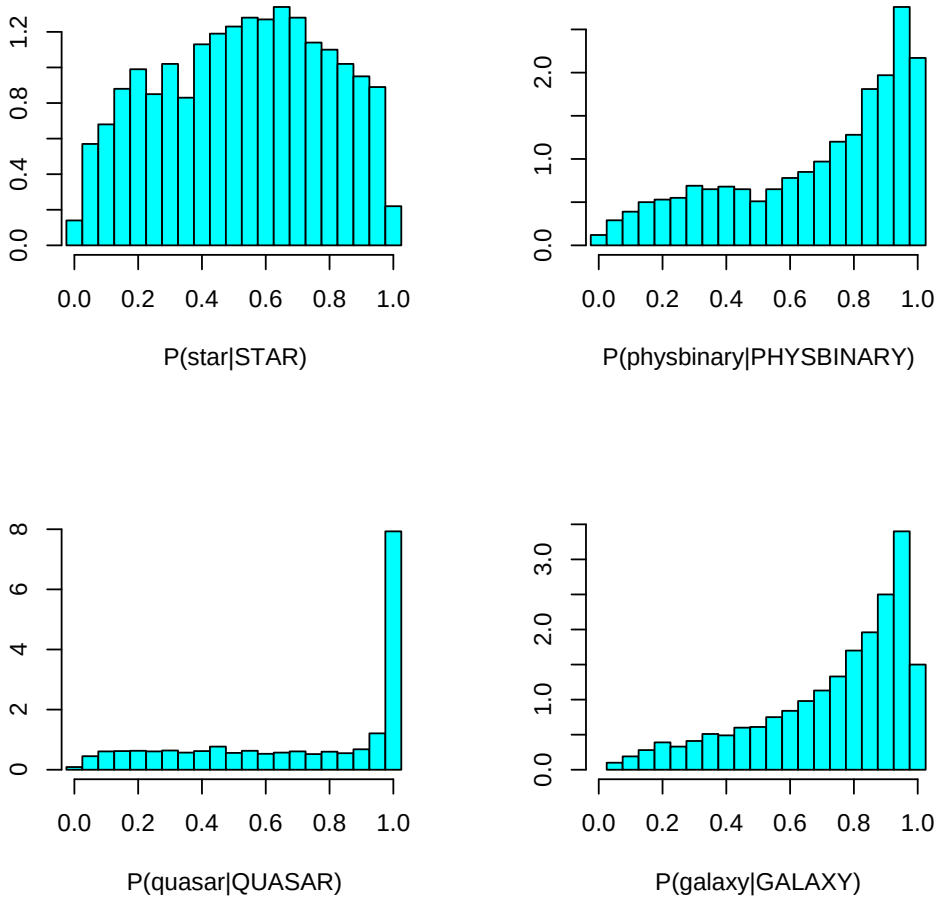


Figure 30: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Nominal priors, G=20

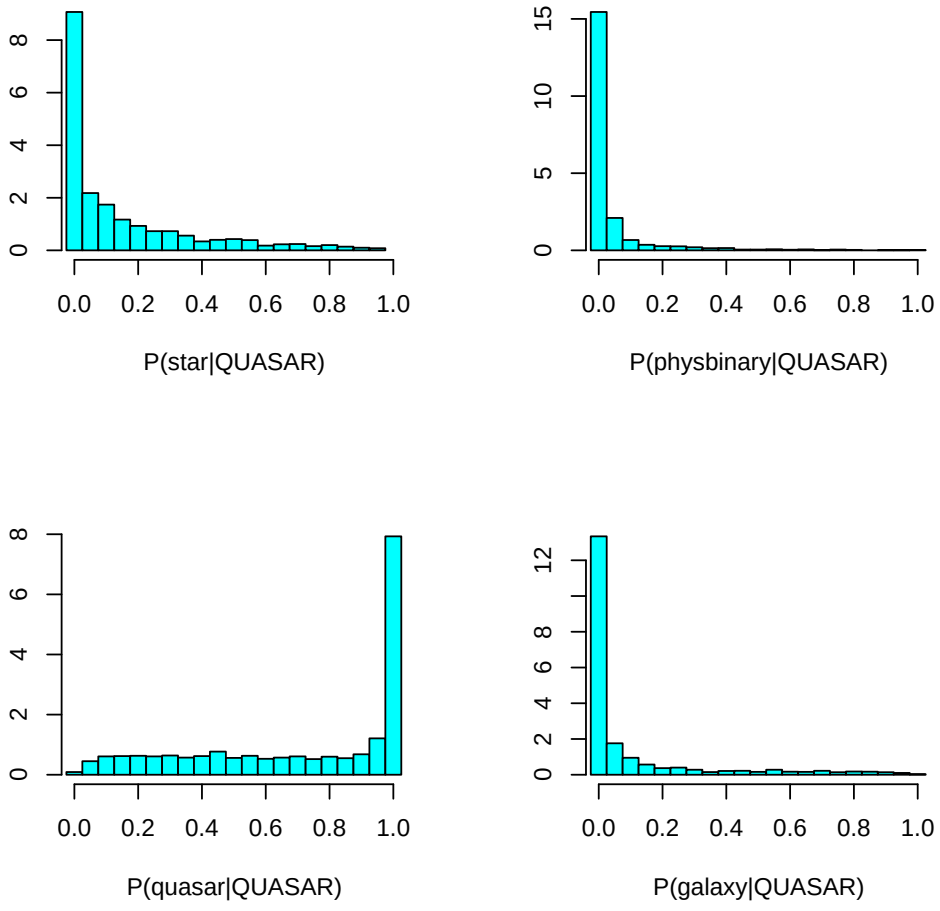


Figure 31: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Nominal priors, G=20

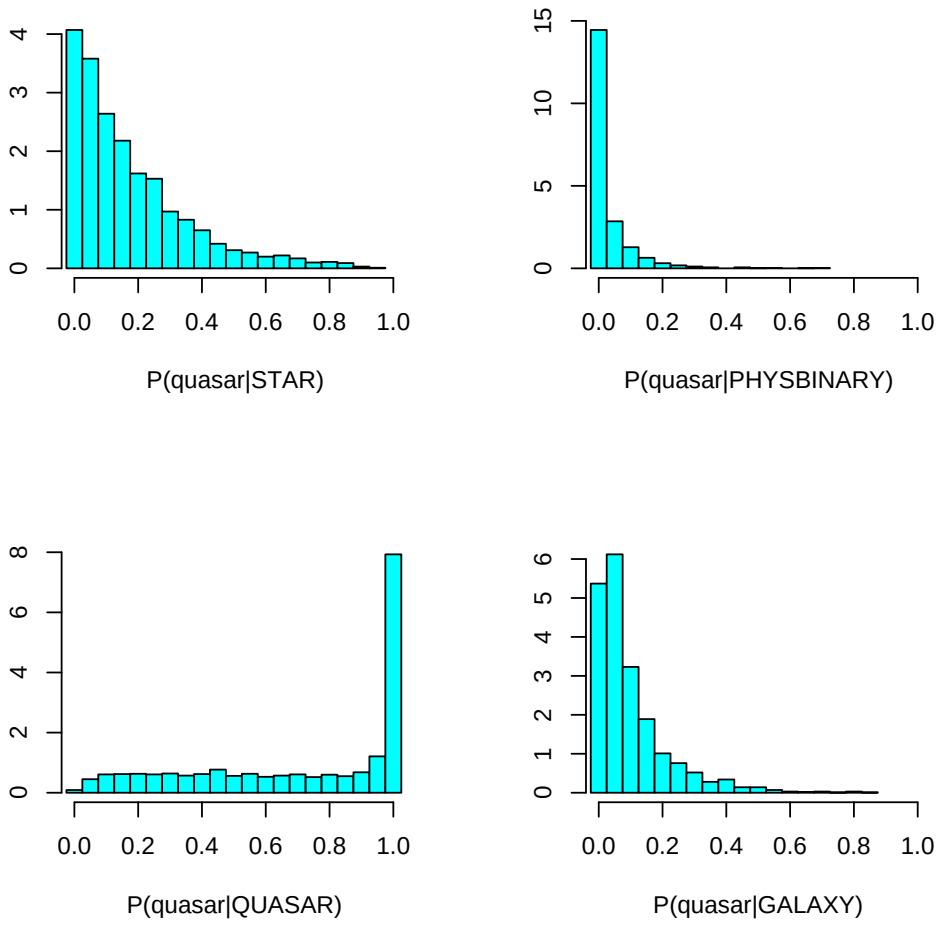


Figure 32: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Nominal priors, G=20

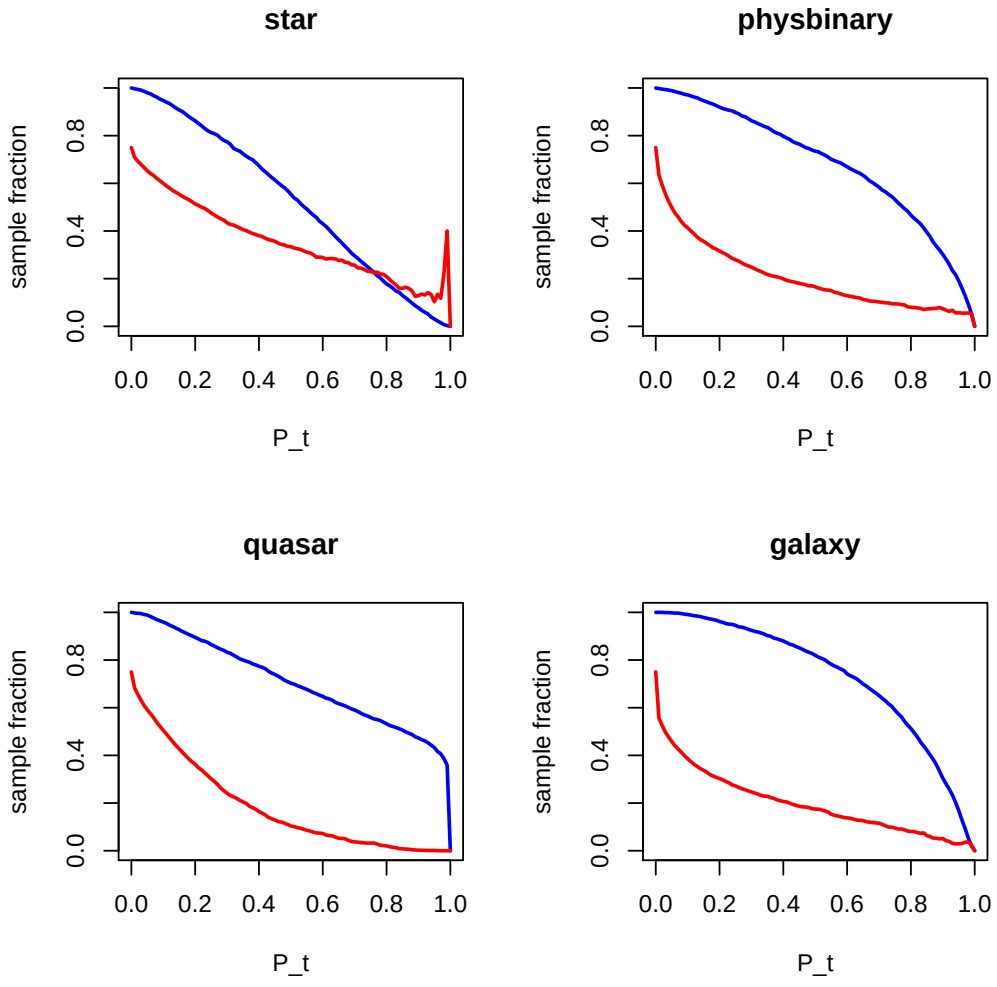


Figure 33: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Nominal priors, G=20

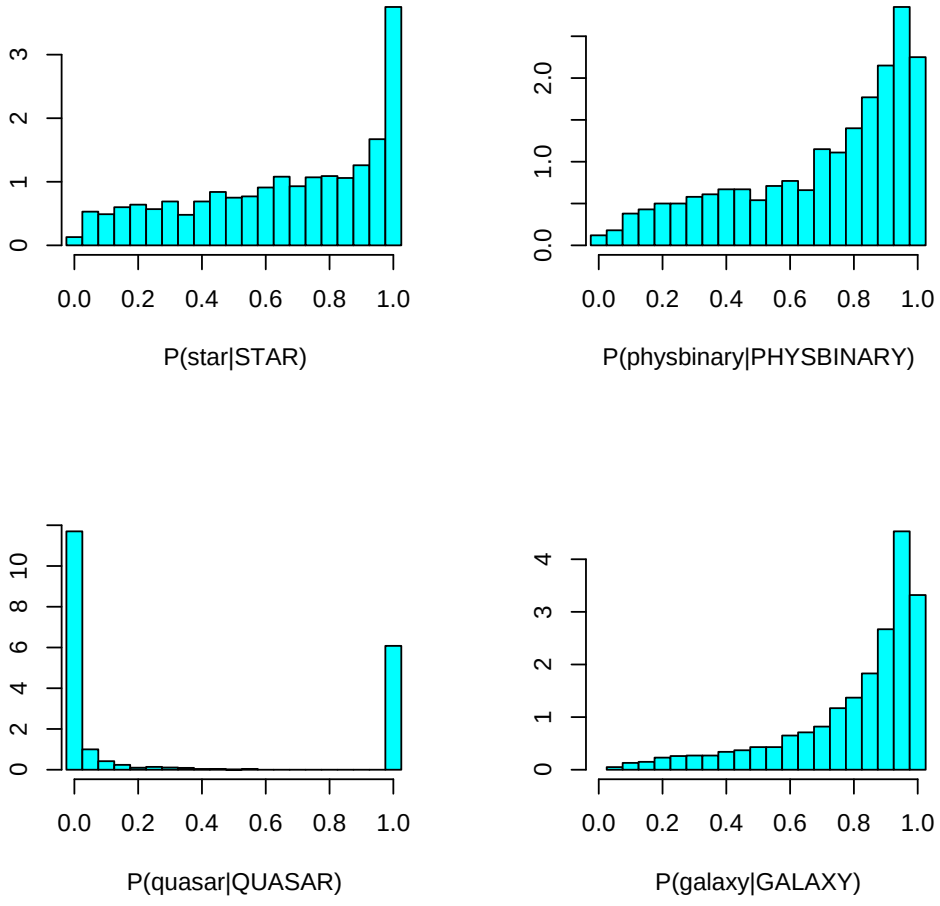


Figure 34: Histograms of DSC outputs for each class showing how confident DSC is of identifying each class. Modified priors, G=20

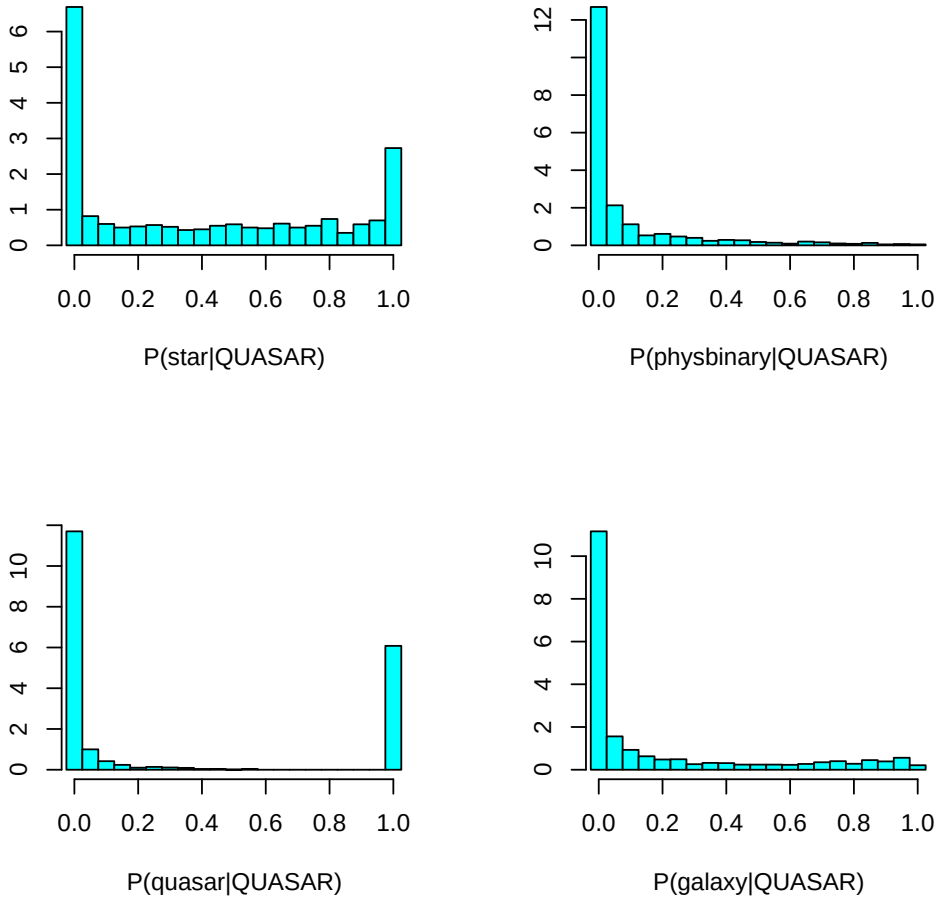


Figure 35: Histograms of DSC outputs showing the probabilities assigned to sources which are truely quasars for each output class. Modified priors, G=20

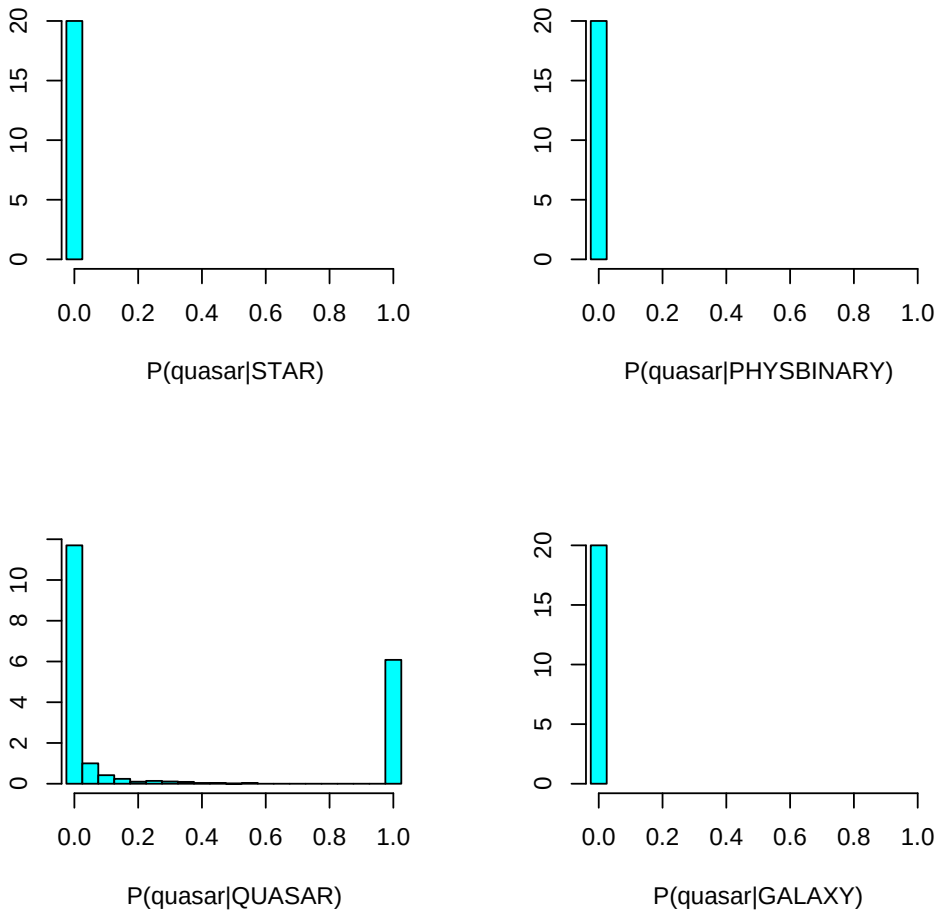


Figure 36: Histograms of DSC quasar output showing the distribution of probabilities for objects of each true class. Modified priors, G=20

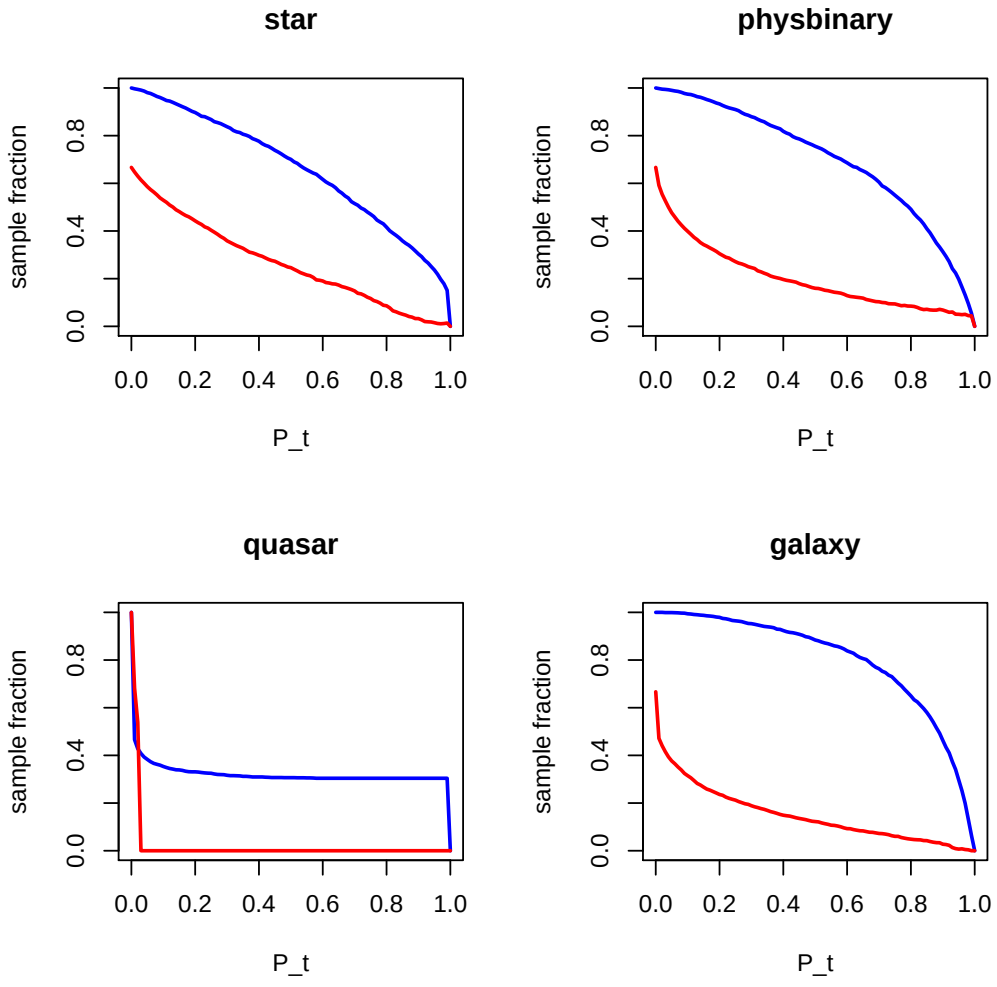


Figure 37: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Modified priors, G=20

# 7 Discussion

## 7.1 The nominal case

The nominal case refers to the unmodified output probabilities. At G=15 we get excellent accuracy with high confidence classifications. Highly complete samples of all classes can be formed with minimal contamination. The biggest confusions is between single stars and physical binaries, as is clear from the plots, but this is still less than 7% (Table 1). Recall that for the contingency tables objects are classified simply by taking the largest probability. So if the probability of an object is above 0.5 for some class, it must be classified as that class. Furthermore, the probability could be almost as low as 0.25 and it still be classified as that class.

At G=18.5 the performance has degraded: some 13% of true stars are misclassified as physical binaries, and, it so happens, vice versa. The probability histograms show that the confidence for true stars and physical binaries (Fig. 14) is much worse than at G=15. This is the impact of noise. Confidence in the extragalactic objects remains good and galaxy classification is little degraded. There is an increased tendency for quasars to be misclassified as stars (7.3%) and vice versa (6.7%) (Table 2), apparent also from the histogram of $P(\texttt{star}|\texttt{QUASAR})$ (Fig. 15). This results in increased contamination and reduced completeness of stellar samples (compare the top-left panels of Figs. 9 and 17). Closer inspection of the data reveals that the quasar contamination drops to zero at $P_t = 0.98$, at which point the completeness is 52%.

It should be realised that single stars and physical binaries are intrinsically difficult to distinguish based only on their spectra. After all, the binaries are simply composite spectra of single stars and if the brightness ratio is large the secondary will make very little contribution to the observed spectrum.

If we now leave out astrometry from the classifiers (at G=18.5), the main impact is lower confidence for stars and physical binaries (top two panels of Fig. 22) and increased confusion between them (Table 3). This is probably because of their very astrometry distributions. We saw in Fig. 5) that physical binaries typically have larger parallaxes and (especially) proper motions. Thus the astrometry alone therefore gives some discriminatory power. This is unrealistic (and ws unintentional in the simulations), but as the performance with the extragalactic classes is essentially unaffected, it probably does not impact the main study of this paper.

Moving from G=18.5 to G=20 (both with astrometry included) the performance degrades further. As we would expect, there is increased confusion between classes. Interesting is the almost flat distribution over $P(\texttt{star}|\texttt{STAR})$ in Fig. 30. This results in signifcant confusion between single stars and physical binaries and difficulty in getting pure or complete samples. Comparing to the case at G=18.5 (Fig. 14), at G=20 we see many more lower probabilities amongst the galaxies, an effect of noise. The quasars, however, are still holding up well. True quasars are still assigned high quasar probabilities (if it's a quasar, we are likely to find it) and mostly low

non-quasar probabilities (Fig. 31), with the main source of "loss" being to the stars (top-left panel). Fig. 32 shows that the main contaminant among quasars is (single) stars followed by galaxies. The fact that physical binaries are a less significant contaminant than single stars is a consequence of larger average astrometry for the former.

## 7.2 The modified case: Quasars 1000 times rarer (G=18.5)

Examining the probabilities, the results agree with what we would expect. Whereas the nominal model identified quasars with high confidence (output probability), it is now much less confident and misses many of them (see Fig. 18 and compare with Fig. 14). In the nominal case the classifier had high confidence for a lot of true quasars, resulting in a large peak in the highest bin for $P(\texttt{quasar}|\texttt{QUASAR})$ (0.95–1.00). In the modified case we now see an additional peak in the other extreme bin, 0.00–0.05. This is simply a result of the probability remapping performed by the priors modification (equation 4), whereby the impact of the normalization should not be ignored.

The changed probabilities have a significant effect on the completeness and contamination curves for quasars: compare Fig. 21 to Fig. 17. The quasar completeness now drops very rapidly, levelling of at around 0.4 before dropping again only when the sample size itself drops to zero. The contamiation curve drops equally rapidly, in fact to zero at $P_t = 0.05$. These reflects the probability remapping apparent in the histograms. The class fraction modification produces a more subtle change in the star sample completeness, which now decreases less rapidly with increasing $P_t$. This is a consequence of the now higher confidence for true stars (compare $P(\texttt{star}|\texttt{STAR})$ in Figs. 14 and 18). In the nominal case a significant number of stars were being assigned relatively high quasar probabilities (top-left panel of Fig. 16). By making quasars a priori rarer we will often assign lower quasar probabilities for *all* objects, and thus higher probabilities to the other classes. Hence the star confidences, $P(\texttt{star}|\texttt{STAR})$, will increase. The star sample is also less contaminated by quasars: by comparing the bottom-left panels of Figs. 15 and 19, we see that in the modified case quasars are assigned lower probabilities of being stars.

The shape of the contamination curve is very convenient for selecting a clean quasar sample. If we take a quasar threshold probability of about 0.1 in Fig. 21, then we will identify about 40% of all quasars with zero contamination. The price we pay for this purity is that it is not possible (with this SVM model and training data) to achieve a higher completeness: as soon as we set the completeness much above 50% we get at least 50% contamination in the sample, which is useless. However, the primary goal is to achieve a clean quasar sample, and a perfect sample of just under half of all quasars is more than sufficient. We of course cannot say that the contamination will remain zero with a larger data set. Instead we have an upper limit on the contamination of 0.5 in 6000 (the number of non-quasars in the test set) which is about 1 in $10^4$.

At first it seems odd that quasars being rarer means that we can achieve a cleaner sample. What's happening is that – as already mentioned – the prior is causing the SVM output probabilities

from the nominal model to be remapped (by equation 4) to allow more discrimination between objects which, in the nominal case, are all almost assiged an output probability of 1.0.

## 7.3 How do we interpret posterior probabilities?

The posterior class probability for an object, equation 2, reflects our belief or degree of certainty that the object is of that class. It is based on both evidence from the data (the likelihood) and on the prior. Specifically, the posterior is a product of these two numbers, normalized across the (four) classes (the normalization if provided by the denominator in equation 2). Can we use these individual posteriors to learn something about the sample as a whole? Consider that we construct a sample of quasars by setting $P_t = 0.3$. For a concrete example, consider the G=18.5 data in the modified case (section 6.2). This comprises 872 objects. All posterior probabilities will be larger than 0.3 (by construction) and in this example the mean probability is 0.973. We may be tempted to make a statement like "we expect 97.3% of the objects to be true quasars", i.e. for a sample of 872 objects, $(1-0.973)*872 = 24$ would not be quasars. This is incorrect, because the posterior probabilities are not statements about the frequency distribution of the *sample*.[12] First, we have used priors which are very non-uniform, namely class fractions of $(1,1,0.001,1)$, and posteriors are not frequencies. Second, and more fundamentally, the set of posterior probabilities are a set of independent degrees of belief on individual objects: A histogram of $P(\texttt{quasar})$[13] is not a probability distribution over some single posterior probability for the sample.

How can we make probabilistic statements about the sample as a whole? If we had a sample size of two with $P(\texttt{quasar}) = \{p_1, p_2\}$ then the probability that at least one object is a quasar is $1-(1-p_1)(1-p_2)$, and we can derive similar (albeit more complex) results for larger samples. If we take the simpler case in which $P(\texttt{quasar}) = p$ for all objects in a sample of size $n$, then the probability that exactly $r$ of them ($r \leq n$) are non-quasars is given by the well known formula

$$P(n,r,p) = \frac{n!}{r!(n-r)!}(1-p)^r p^{(n-r)} \tag{7}$$

We could use this to make approximations for the above sample ($n = 872$) by setting $p = 0.973$, the mean posterior probability. With $r = 0$, we see that the probability that no object in the sample is a non-quasar is $10^{-11}$, i.e. we are almost certain to have at least one contaminant. The fact that we didn't is partly because this approximation is poor: 798 of the 872 posterior probabilities are equal to or greater than 0.999. Using this as an alternative approximation we have $P(n = 872, r = 0, p = 0.999) = 0.42$; there is a 0.58 chance we'll have at least one contaminant. The left panel of Fig. 38 shows how $log_{10}P(n = 872, r, p = 0.973)$ varies with $r$. (In case one is interested it has a maximum of $P(n,r,p) = 0.082$ at $r = 23$). Summing equation 7 over ranges of $r$ we can make a more useful statement. For example, summing from $r = 0$ to

[12] Incidentally there is not a single contaminant in this sample: $P_t$ was chosen from inspection of Fig. 21 to ensure this. (Yet this does not imply that the probability that all 872 objects are quasars is unity!)

[13] This histogram is the sum of the four panels in Fig. 20 for $P_t > 0.3$.

$r = 24$ (with $p = 0.973$), we deduce that the probability of having 24 contaminants or fewer is 0.56. For 40 or fewer it is 0.9992, i.e. there is a probability of only 0.008 that there are more than 40 contaminants. The right hand panel of Fig. 38 shows this for variable $r$: specifcally it plots

$$1 - \sum_{r'=0}^{r'=r} P(n, r', p) \tag{8}$$

with $n = 872$ and $p = 0.973$ against $r/n$. This is the probability that the contamination fraction of the sample is more than $r/n$. Thus at least in the case where we approximate the posterior probabilities to be equal we can make probabilistic inferences about the contamination of the sample. With a test data set we could compare these with the actual contamination level. [14]

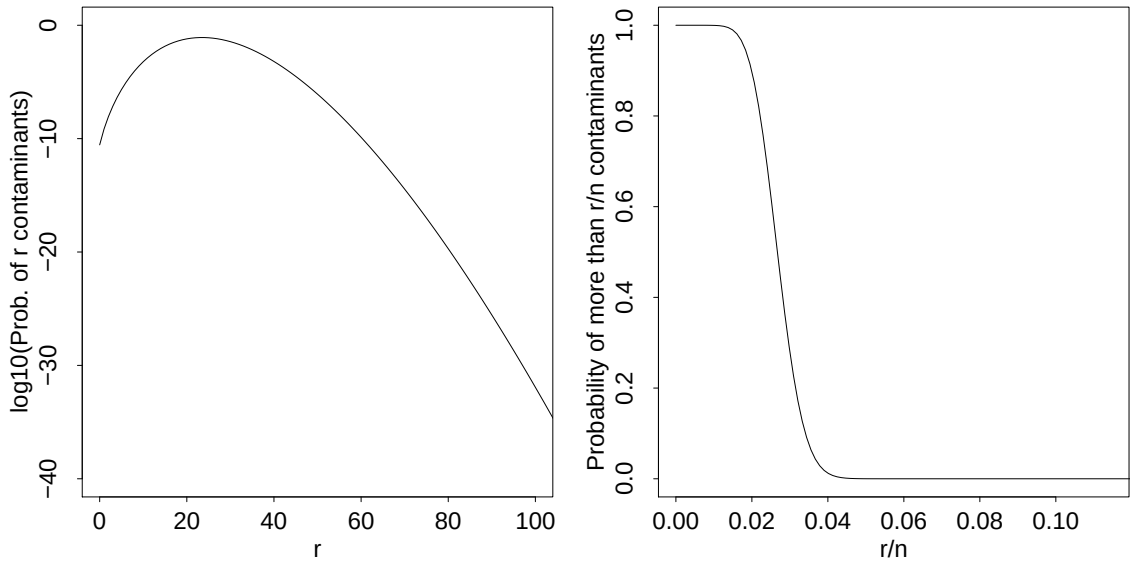


Figure 38: The left panel shows the variation with $r$ of $P(n = 872, r, p = 0.973)$ (equation 7), the probability that there are exactly $r$ contaminants in the quasar sample of $n$ objects, each with posterior quasar probability of $p$. The right hand panel is 1 minus the cumulation of this (equation 8), and so shows the probability that there will be more than $r$ contaminants (shown as the contamination fraction $r/n$).

The take-home message of this discussion is that the DSC posterior probabilities are statements, whether in the nominal or modified case, about *individual* objects. To make inferences about the *sample* you must be careful to make logically consistent combinations of the individual probabilities.

[14] Curiously the curve in Fig. 38 drops to zero at $r/n = 0.04$, close to the value of $P_t$ at which the contamination curve drops to zero.

## 7.4 Testing the predictions

We saw in section 7.2 how the modified model, with modified class fractions of (1,1,0.0001,1), resulted in a a dramatic change in the completeness and contamination curves. By choosing an appropriate threshold the contamination could be greatly reduced, at the cost of also reducing the completeness.

What happens if we repeat this but with a different set of modified class fractions, say (1,1,0.1,1), i.e. with quasars just ten times rarer? The resulting plot for the G=18.5 with astrometry data is shown in Fig. 39. Comparing this with the nominal case (Fig. 17) and the original modified case (Fig. 21), we see that the curves for the quasars in the new case are intermediate, as we would expect. For example, the contamination now only drops to zero at $P_t \geq 0.83$, at which level the completeness is 52%. So by selecting an appropriate threshold we can achieve the same levels of completeness and contamination as before.

All of the completeness and contamination curves shown so far for the modified model have been calculated using the nominal test data set but with the modified equations (equation 5). They are, therefore, just predictions of how well the modified model *would* perform *if* it were applied to a test data set which really had the modified class fractions. We now test the accuracy of these predictions by using a new test data set which really does have class fractions (1,1,0.1,1). This test set is built by randomly removing 90% of the quasars from the original test set. (The output probabilities are calculated with the modified model, equation 4, but the completeness and contamination must of course be evaluated with equation 1.) This is shown in Fig. 40. Comparing the quasar curves with those in Fig. 39, we see that the agreement is very good, especially for the contamination. A better comparison is possible with Fig. 41. The discrepancy between predicted and actual contamination varies from 0–10%, averaging 5%.

## 7.5 A final word

Ultimately we would like to know whether the modified model is "better" than the nominal one, i.e. does the modification of the priors bring advantages. If we ignored the issue of priors and class fractions, then our predictions of completeness and contamination would be those from the nominal model (equal class fractions) in section 6, e.g. Fig. 17. Yet the actual completeness and contamination we would measure in a situation with difference class fractions, say (1,1,0.1,1) are quite different, namely those shown in Fig. 42. So our conclusions with this approach would be wrong.

If we are only interested in a clean quasar sample, then it turns out that the predictions from the nominal model are good. We saw from Fig. 17 that we could form a zero contamination sample with completeness of about 50%, just by setting a much larger threshold ($P_t = 0.98$) than with the modified model. Looking at the results on the test data set with class fractions (1,1,0.1,1) this threshold also gives zero contamination and a completeness of 55% (Fig. 42). So at least in

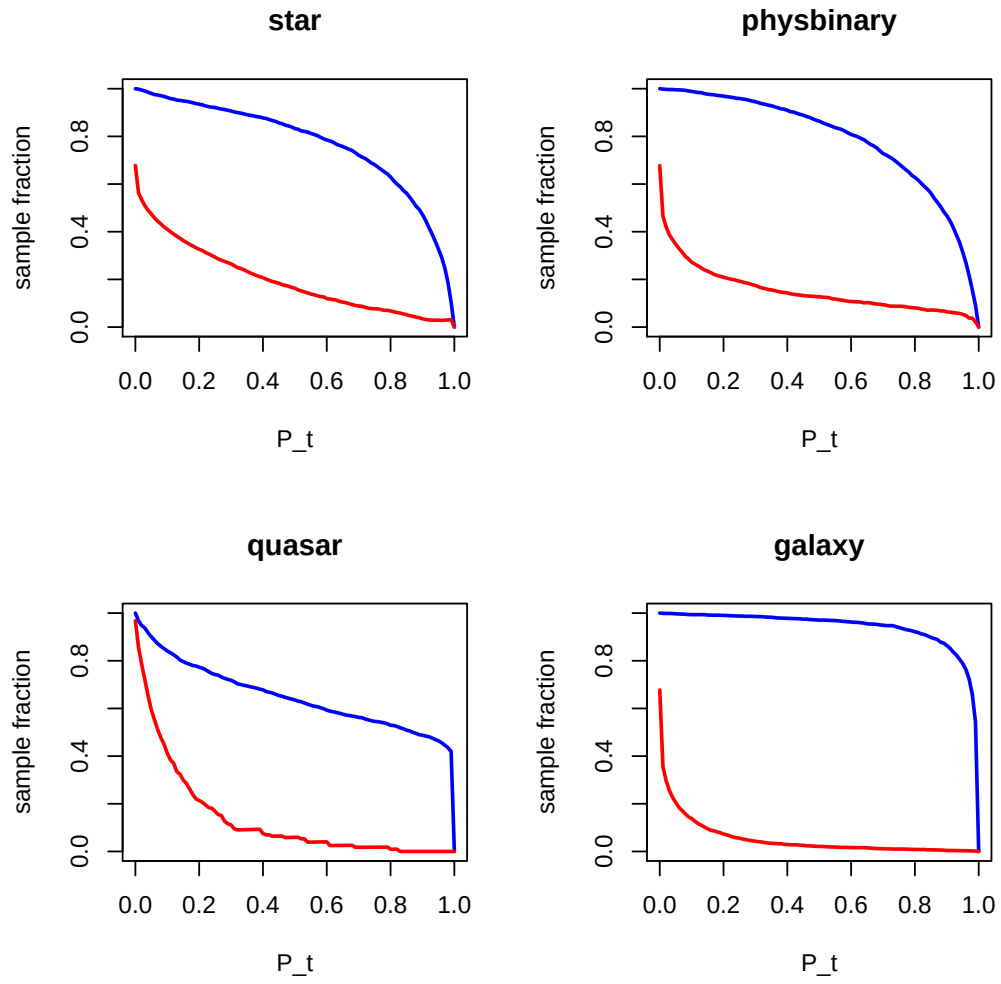


Figure 39: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. Now using modified priors with class fraction (1,1,0.1,1). G=18.5, with astrometry. As before, we use this to predict C&C.

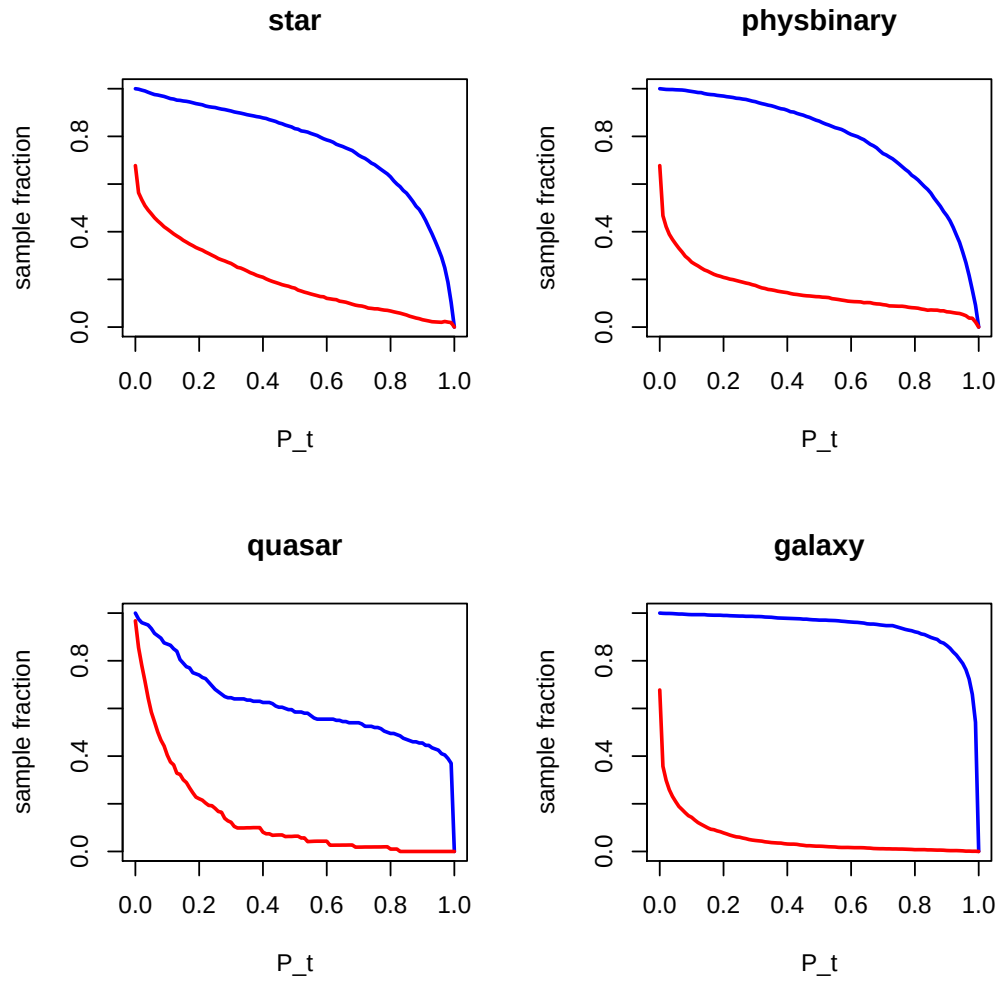


Figure 40: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. This uses the modified model with class fractions (1,1,0.1,1) to determine the C&C in a new test sample which has these actual class fractions.

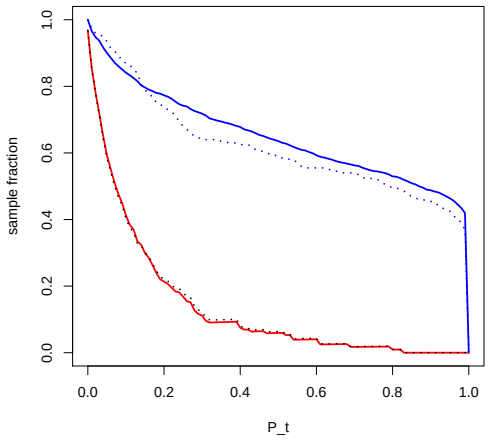


Figure 41: Direct comparison of the predicted and measured quasar completeness and contamination. The solid lines are the predictions from Fig. 39. The dashed lines are the measured from Fig. 40, whereby the contamination is plotted in black rather than red to aid distinction.

this case, the threshold derived from the nominal model is accurate.

We may argue that actually Fig. 42 should be our predictions from the nominal model, i.e. we just modify the test data set but not the model itself. How then, can we make a comparative test of the nominal and modified models and decide which to use? Is not yet clear (to us) how we test the accuracy of a probabilistic model when applying a threshold. We cannot define a single accuracy figure, because the accuracy (or rather, completeness and contamination) vary with the threshold chosen to form the sample, and we just saw that we would set different thresholds in the two cases. It remains to be tested whether the two methods give equally complete samples for a variable contamination (or vice versa) for different class fractions and sample sizes.[15] We also need to study whether one approach gives more self-consistent posterior probabilities for the sample, an issue we started to discuss in section 7.3.

# 8 Summary and Conclusions

- To build a clean sample of quasars we need to use a probabilistic classifier and only select as quasars objects which have a probability above some threshold. We show how we can control the sample completeness and contamination by varying this threshold.

[15] It may turn out that for building samples it is sufficient to use the nominal model and to calculate the completeness and contamination using the modified class fractions (equation 5) in order to select the appropriate threshold.

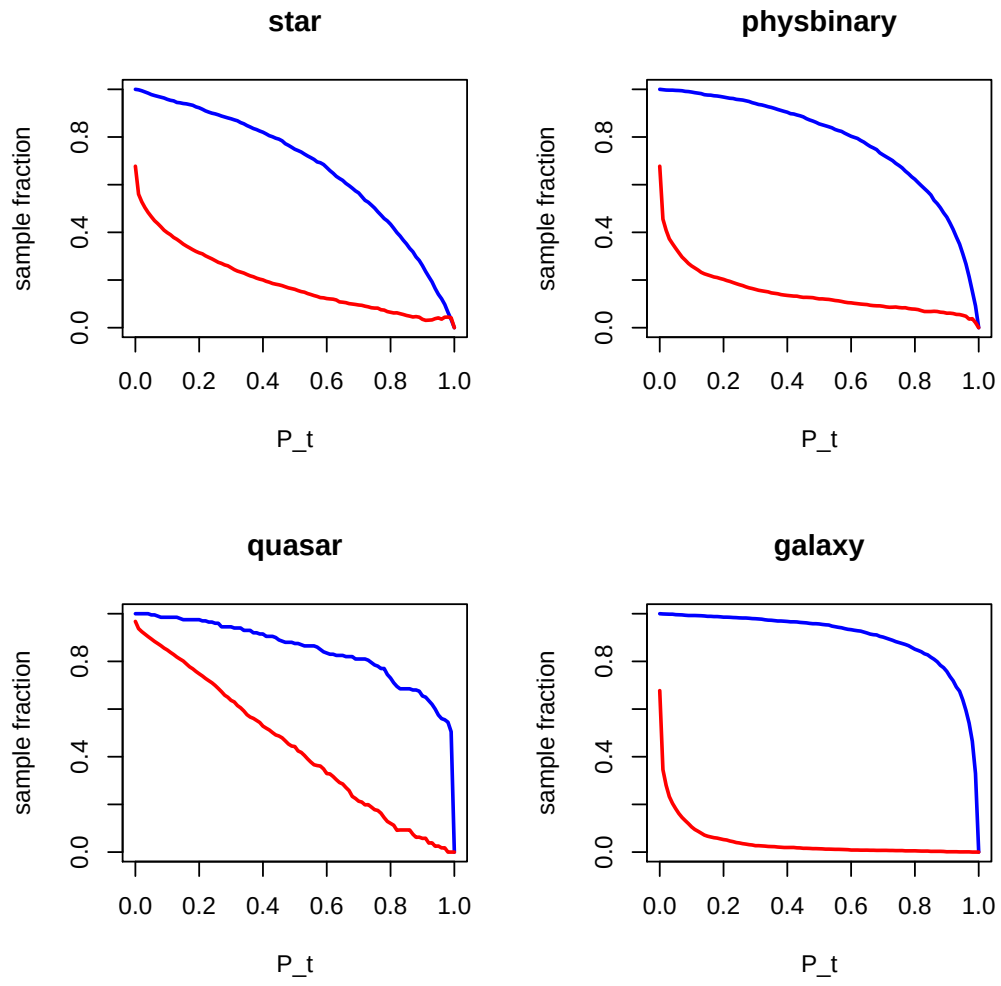


Figure 42: Completeness (blue/upper line) and contamination (red/lower line) of a sample as function of the probability threshold. This applies the nominal model to a test data set which has class fractions (1,1,0.1,1).

- With the present DSC software, trained and tested on synthetic BP/RP data for a four-class problem, we can achieve a contamination of less than 1 in $10^4$ in the QSO sample which is 51% complete at G=18.5 or 43% complete at G=20, for the case that the relative fractions of quasars to stars is 1/1000. Including astrometry in the input data does not significantly improve this. This upper limit on the contamination is set by the sample size. We can determine whether it can be decreased by applying the model to a larger test data set (future work).

- The class priors are often not explicit in a classification model, and sometimes are implicitly influenced by the training data distribution. We have developed a way of recalculating the DSC posterior probabilities to reflect a user-defined prior, without having to change the training data or retrain the model. The prior is our pre-data estimate of the probability that the object has a certain class. (It might be based on the Galactic latitude and G magnitude of the object, for example.) Using the nominal model, i.e. that trained on equal class fractions, we can remap its output (posterior) probabilities to be relevant to different class fractions (this defines the *modified model*). We apply this to a situation in which quasars are 1000 times rarer than stars (our prior). This flexibilty to set the prior allows us to build a classification model (the modified model) which is independent of the class frequencies in the training data.

- We have introduced a simple method for calculating the implicit priors in a classification model, what we call the *model-based priors*. We show that these priors agree with the true class fractions in the training data set in the nominal case and with the effective training class fractions in the modified case. This ensures a self-consistency of the method for building the modified model.

- We demonstrated that the predicted completeness and contamination for the modified model are accurate. Specifically, they agree well with the actual completeness and contamination calculated from a test data set which has class fractions equal to those assumed by the model (the model-based priors).

- We show that ignoring priors (as expressed via the class fractions) will lead to wildly incorrect estimates of sample completeness and contamination if the model is applied to data with class fractions which are very different from that assumed in the model.

# 9 Acknowledgements

We thank Carola Tiede for useful discussions on probabilities and classification and Sebastian Jester for insight into the SDSS quasar and star sample biases.

# 10 References

CBJ-029, Bailer-Jones 2007, Cycle 3 simulation specifications, on Livelink (CU8 folder)
CBJ-037, Bailer-Jones & Smith 2008, Multistage classification for DSC, in preparation
KS-003, -004, -005, Smith 2007, Cycle 3 DSC documents, Gaia technical notes
http://gaia.esac.esa.int/dpacsvn/DPAC/CU8/docs/cycle03/DSC/
Richards et al., 2002, AJ 123, 2945
Richards et al., 2004, ApJSS 155, 257
Richards et al., 2006, AJ 131, 2766
RS-002, Sordo & Vallenari 2008, Cycle 3 simulations description, on Livelink (CU8 folder)
Schneider et al., 2005, AJ 130, 367
Shin & Cho, 2003, Proc. of the Korean Data Mining Conference, 93
http://www.kyb.mpg.de/publication.html?publ=2709
Van den Berk et al., 2005, ApJ 129, 2047
Visa & Ralescu, 2005, Proceedings of the Sixteenth Midwest Artificial Intelligence and Cognitive Science Conference, 67
http://www.ececs.uc.edu/~aralescu/PAPERS/VRMaics2005.pdf
Weiss, 2004, ACM SIGKDD Explorations Newsletter, 6, 7
http://portal.acm.org/citation.cfm?id=1007734

# A Quasar to star ratio, completeness and contamination in SDSS

Quasars are rarer than stars, a fact revealed by many surveys. Here we use SDSS DR6 to quantify this as a function of magnitude and Galactic latitude.

SDSS initially observes the sky in the five ugriz bands. It then makes a selection based on these colours to identify quasar candidates, some of which are then subject to spectroscopic follow up (Schneider et al. 2005, Richards et al. 2002, 2006). These spectra are classified to confirm or refute quasar status. Other objects (in particular stars and galaxies) are also targetted for spectroscopic follow up.

## A.1 Object selection and quasar sample properties

We here use two flags in the SDSS catalogue to count the relative numbers of quasars to stars. These are `PhotoPrimary.type`, which is a morphogical classification, and `SpecObj.specClass`, which is a spectral classification. The relevant values for the former are 3 for galaxy and 6 for star-like objects (including stars and quasars). For the latter, 1 and 6 denote star, 3 and 4 quasar.

Our stellar sample comprises all star-like objects according to the `type` flag, but if they have a spectrum this must be a stellar spectrum. Our quasars are any star or galaxy photometric object which is spectroscopically confirmed to be a quasar. In this way we include quasars with a visible galaxy. The SQL queries to achieve these samples are

**Quasar selection**

```
select round(PhotoPrimary.l,2) as l, round(PhotoPrimary.b,2) as b,
  round(PhotoPrimary.g,3) as g, round(PhotoPrimary.i,3) as i,
  PhotoPrimary.type, SpecObj.specClass
into mydb.test18
from PhotoPrimary
left join SpecObj
on PhotoPrimary.specObjID=SpecObj.specObjID
where
    PhotoPrimary.i between 0.0 and 22.0
    and PhotoPrimary.type in (3,6)
    and SpecObj.specClass in (3,4)
```

**Star selection**

```
select round(PhotoPrimary.l,2) as l, round(PhotoPrimary.b,2) as b,
  round(PhotoPrimary.g,3) as g, round(PhotoPrimary.i,3) as i,
  PhotoPrimary.type, SpecObj.specClass
into mydb.STAR1055
```

```
from PhotoPrimary
left join SpecObj
on PhotoPrimary.specObjID=SpecObj.specObjID
where
    PhotoPrimary.l between 54 and 56
    and PhotoPrimary.b>=0
    and PhotoPrimary.i between 0.0 and 22.0
    and PhotoPrimary.type=6
    and ( (PhotoPrimary.specObjID=0) or
      (PhotoPrimary.specObjID!=0  and SpecObj.specClass in (1,6)) )
```

`SpecObjID` is used to track the existence of a spectrum. We select stars in four longitude strips each 2 degrees wide in longitude centered at $b = 55, 65, 195, 205$. These are chosen because they extend to the lowest Galactic latitudes (about $b = 20$).

We then calculate, for each longitude strip and for a fixed latitude range, the number of stars and the number of quasars in 1-mag. wide magnitude bins. In total we find 87 525 quasars in the range $i = 0, 22$ in the 9583 sq. deg. in DR6, and 85 438 in the range $i = 0, 20.2$ an average of 8.9 quasars per sq. deg. This compares to 46 420 in the SDSS DR3 quasar catalogue of Schneider et al. (2005) over 4188 sq. deg. (11.1 per sq. deg.) down to $i = 20.2$. A better comparison may be with the sample Richards et al. (2006) used to construct a $z$ =0–5 quasar luminosity function sample (and is itself derived from the Schneider et al. sample). This gives 15 343 over an effective area of 1622 sq. deg. (9.5 per sq. deg.) and extends from $i = 15$ to $i = 19.1$ for $z < 3$ and to $i = 20.2$ for $z > 3$. Fig. 43 and 44 show the $i$ and $g$ band magnitude distribution for the QSOs in our sample and Fig. 45 the colour magnitude distribution. The spatial distribution in Galactic coordinates of course follows the SDSS imaging footprint (Fig. 46).

Figs 49 and 50 plot the quasar/star ratio as a function of magnitude for Galactic latitude bins in four Galactic longitude strips.

## A.2 Sample completeness

There are, of course, various selection biases which differentially affect the quasar and star number counts, and therefore the derived quasar/star ratio. According to Vanden Berk et al. (2005), the SDSS QSO survey is surprisingly complete. By taking spectra of all point sources in 233 sq. deg. region down to a redenning corrected limit of $i = 19.1$ (the stripe 82 completeness survey), they find that the completeness of the colour-selected, unresolved quasar sample is about 95%. Allowing for the fact that extended quasars are less efficiently detected, the combined "end-to-end" completeness of the SDSS quasar survey is about 89%. 4% of this loss is accounted for by image defects and blends. In addition, the quasar selection has a lower efficiency at redshifts between 2.5 and 3.0 because of the degeneracy (locus crossing) in the stellar and quasar SDSS colours (Richards et al. 2006). The contamination at this level of completeness is 24.7%. This

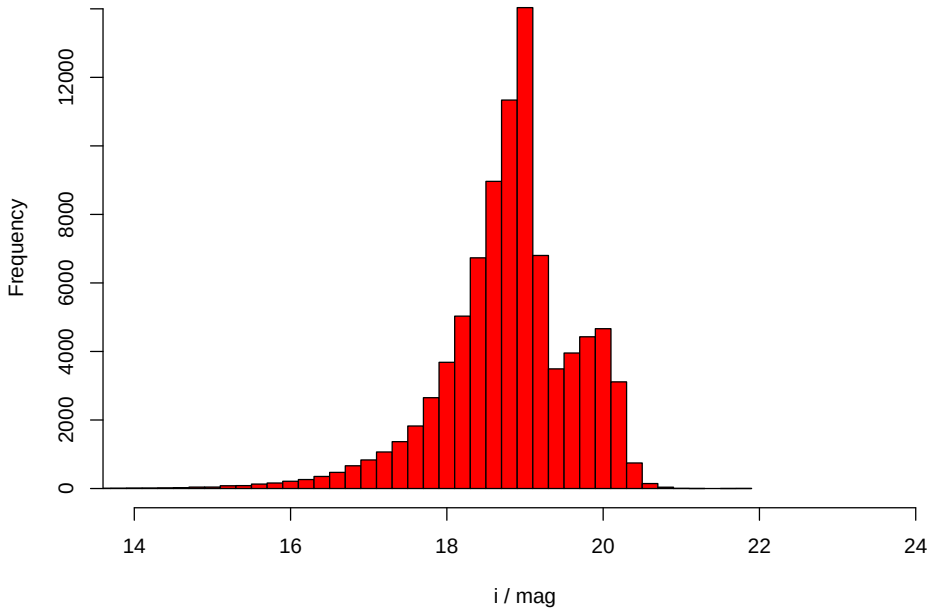


Figure 43: QSO SDSS DR6 i-band magnitude function

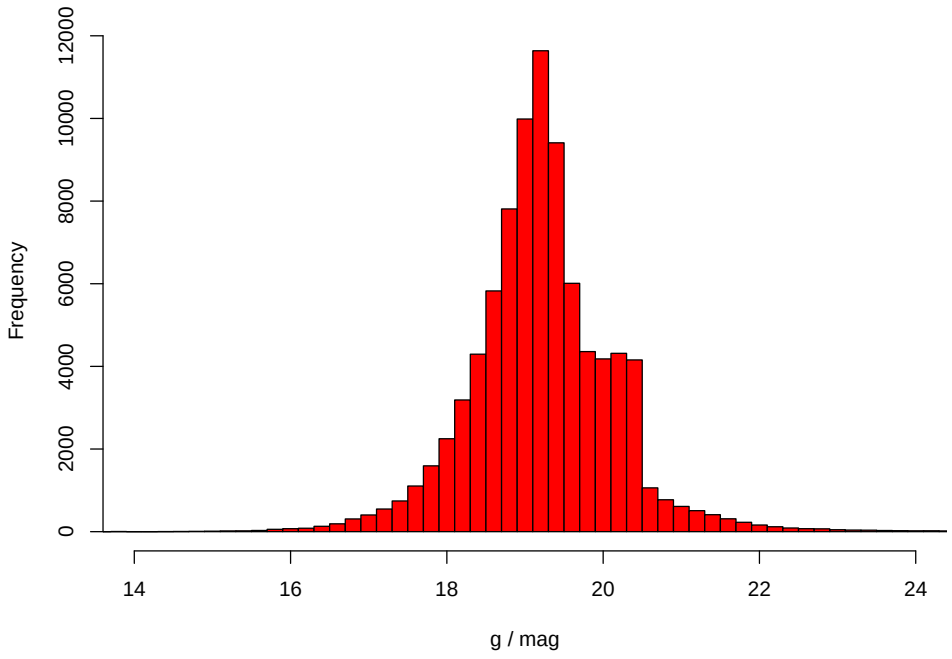


Figure 44: QSO SDSS DR6 g-band magnitude function

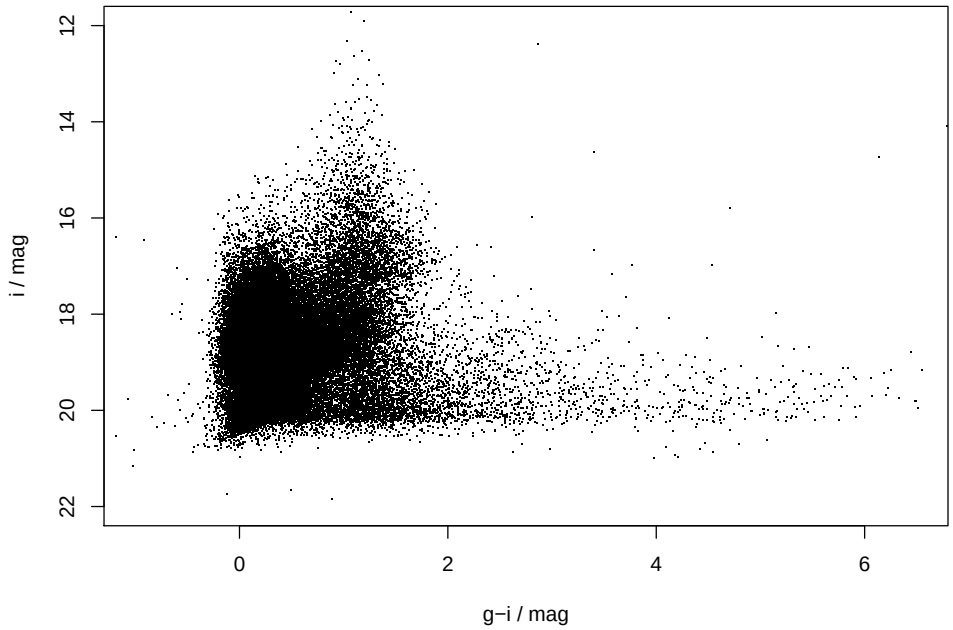


Figure 45: QSO SDSS DR6 colour-magnitude diagram

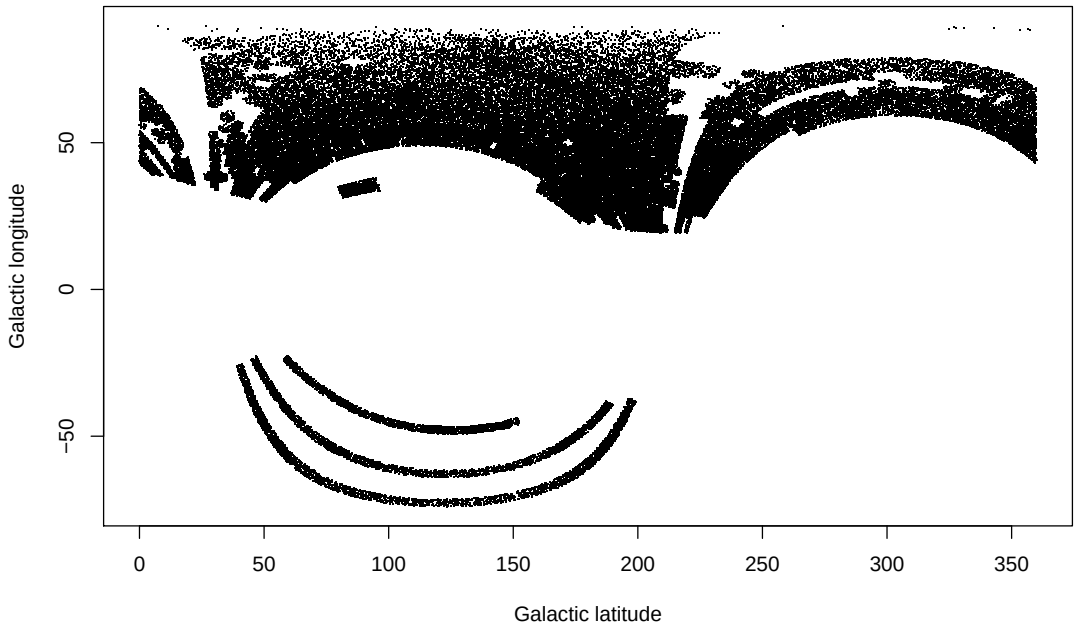


Figure 46: QSO distribution in SDSS DR6

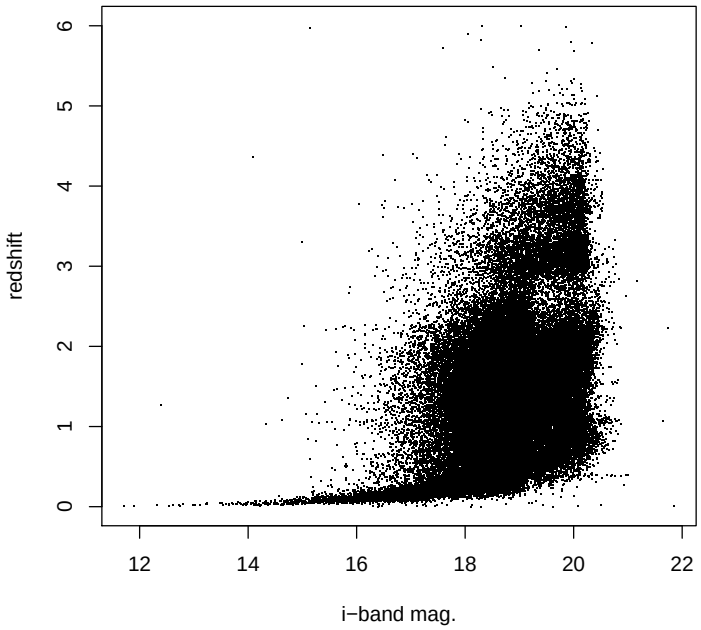


Figure 47: QSO SDSS DR6 i-band, redshift relation

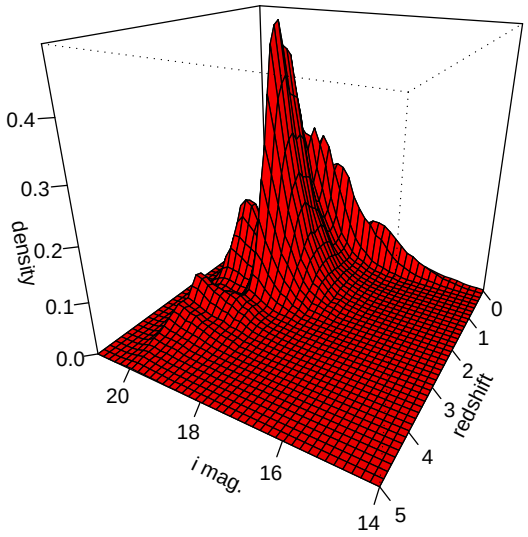


Figure 48: QSO SDSS DR6 i-band, redshift density plot

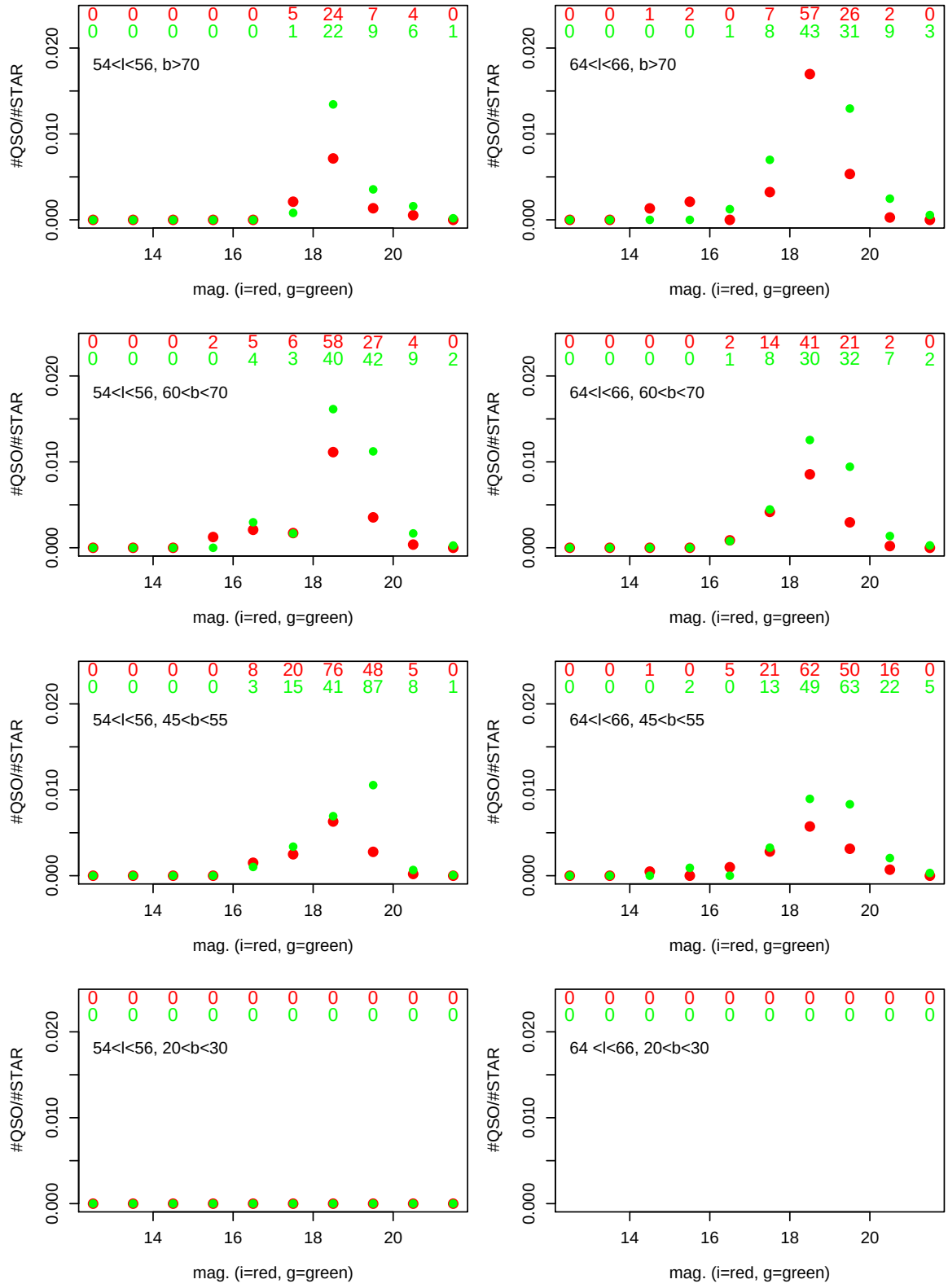


Figure 49: The ratio of QSOs to stars in SDSS as a function of $i$ magntiude (large red dots) and $g$ magnitude (smaller green dots) for different Galactic longitude, $l$, strips (columns) and Galactic latitude, $b$, ranges (rows). The numbers in red (green) along the top of each panel are the absolute numbers of quasars in each bin.

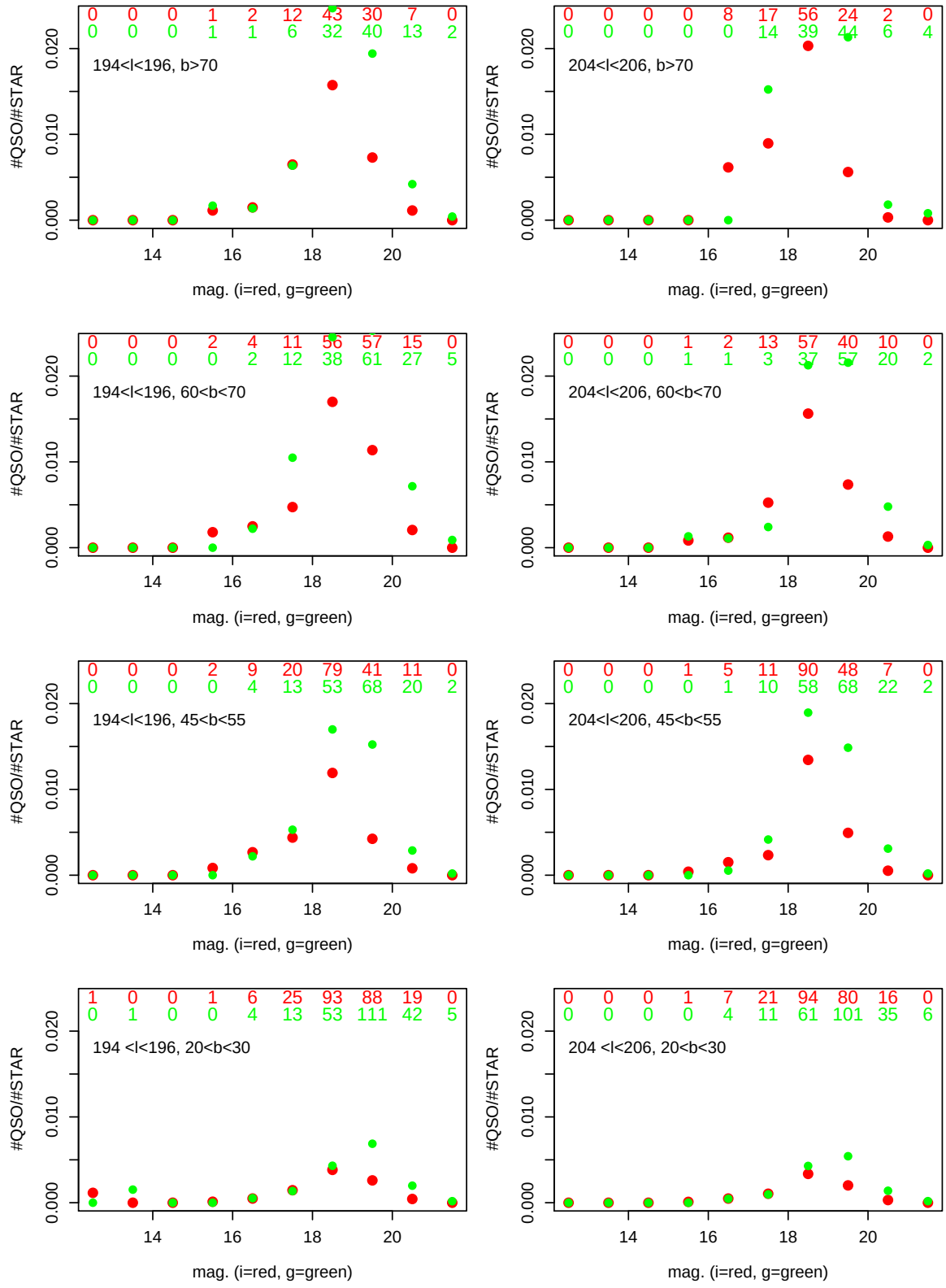


Figure 50: As Fig. 49, but for l=195, 205

compares to 34% from Richards et al. 2002 (where the quasar spectroscopic target selection algorithm is defined), although this includes more high-z quasars where the contamination is higher. At magnitudes fainter than $i = 19.1$ the completeness is lower, and quasar candidates fainter than $i = 20.2$ or brighter than $i = 15$ are not selected for spectroscopic follow up at all. Other approaches appear to yield different figures. By training a classifier on photometry of 22 000 spectroscopically-confirmed quasars, Richards et al. (2004) build a photometrically-selected sample of 100 000 quasar candidates to $g = 21$ from DR1. They estimated that the completeness of unresolved UVX quasars to $g = 19.5$ in this sample was 94.7%, with a contamination of 5% to $g = 21$.

The stellar completeness is probably lower (some details from Sebastian Jester, private communiation). First, SDSS has holes in its photometric sky coverage. There are primarily due to high source density and an inability of the photometric pipeline to deblend sources. These occur predominantly in areas of high stellar density, so stars are probably missed more than quasars. As stars are not the main target of SDSS, most stars do not have spectra taken, so some photometric star classifications may actually be quasars. However, the completeness test of Vanden Berk et al. (2005) implies that this is a small fraction.

Overall, at higher Galactic latitudes, we expect reasonable completeness for both the quasar and stellar samples in the range $i = 15 - 19.1$ and so the quasar/star ratios to be reasonably reliable. At lower Galactic latitudes the picture is less clear, but probably more stars are lost than quasars, leading to the quasar/star ratio being overestimated. Of course, these results have been obtained in the SDSS filtes. To apply to Gaia, we would need to convert the colours. An approximate relation based on the Gaia instrument design is shown in Fig. 51.

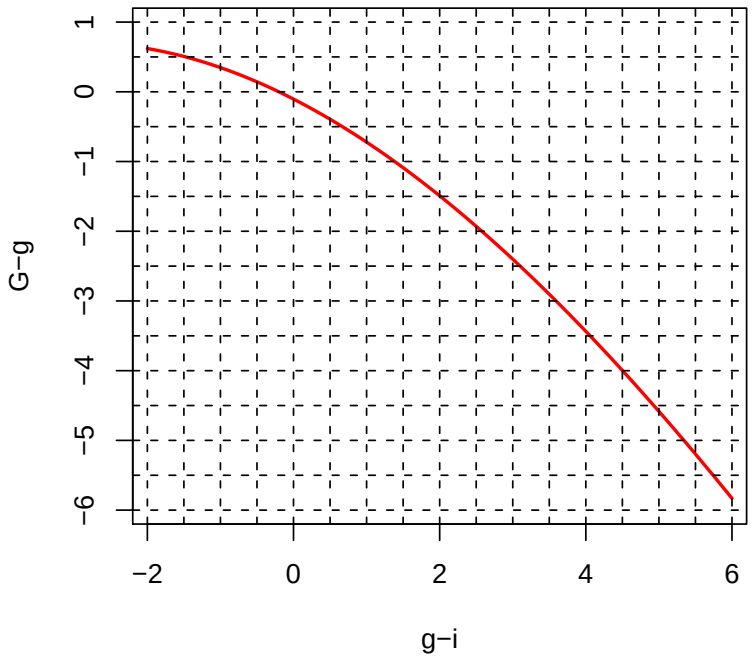


Figure 51: Colour transformation between Gaia G filter and SDSS *g* and *i* filters. From Jordi (2007) GAIA-C5-TN-UB-CJ-041-2.