Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **515,** 3224–3248 (2022) https://doi.org/10.1093/mnras/stac1944
Advance Access publication 2022 July 23

Paving the way for *Euclid* and *JWST* via probabilistic selection of high-redshift quasars

Riccardo Nanni [1,2]★ Joseph F. Hennawi [1,2] Feige Wang,[3] Jinyi Yang,[3] Jan-Torge Schindler[1,4] and Xiaohui Fan[3]

[1]*Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands*
[2]*Department of Physics, University of California, Santa Barbara, CA 93106-9530, USA*
[3]*Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA*
[4]*Max Planck Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*

Accepted 2022 July 6. Received 2022 July 5; in original form 2021 November 5

ABSTRACT

We introduce a probabilistic approach to select $6 \leq z \leq 8$ quasar candidates for spectroscopic follow-up, which is based on density estimation in the high-dimensional space inhabited by the optical and near-infrared photometry. Densities are modelled as Gaussian mixtures with principled accounting of errors using the extreme deconvolution (XD) technique, generalizing an approach successfully used to select lower redshift ($z \leq 3$) quasars. We train the probability density of contaminants on 1902 071 7-d flux measurements from the 1076 deg^2 overlapping area from the Dark Energy Camera Legacy Survey (DECaLS) (z), VIKING ($YJHK_s$), and un*WISE* ($W1W2$) imaging surveys, after requiring they dropout of DECaLS g and r, whereas the distribution of high-z quasars are trained on synthetic model photometry. Extensive simulations based on these density distributions and current estimates of the quasar luminosity function indicate that this method achieves a completeness of $\geq$ 56 per cent and an efficiency of $\geq$ 5 per cent for selecting quasars at $6 < z < 8$ with $J_{AB} < 21.5$. Among the classified sources are 8 known $6 < z < 7$ quasars, of which 2/8 are selected suggesting a completeness $\simeq$ 25 per cent, whereas classifying the 6 known ($J_{\rm AB} < 21.5$) quasars at $z > 7$ from the entire sky, we select 5/6 or a completeness of $\simeq$ 80 per cent. The failure to select the majority of $6 < z < 7$ quasars arises because our quasar density model is based on an empirical quasar spectral energy distribution model that underestimates the scatter in the distribution of fluxes. This new approach to quasar selection paves the way for efficient spectroscopic follow-up of *Euclid* quasar candidates with ground-based telescopes and *James Webb Space Telescope*.

Key words: galaxies: active – quasars: supermassive black holes – early Universe.

1 INTRODUCTION

Luminous high-redshift quasars (QSOs) are amongst the best probes of the primordial Universe at the end of the dark ages. Their spectra provide important information regarding the properties of the intergalactic medium (IGM) during the epoch of reionization (EoR). In fact, deep spectroscopy of $z > 6$ QSOs showed that the IGM is significantly neutral at $z \geq 7$ (e.g. Bañados et al. 2018; Davies et al. 2018; Wang et al. 2020; Yang et al. 2020a), but highly ionized at $z \leq 6$ (e.g. McGreer, Mesinger & Fan 2011; McGreer, Mesinger & D'Odorico 2015; Yang et al. 2020b).

In addition, the engines of the most distant QSOs, the super massive black holes (SMBHs), are crucial for understanding the formation mechanisms of the first generation of black hole seeds (see Inayoshi, Visbal & Haiman 2020, for a recent review). Their existence up to $z = 7.6$ (e.g. Wang et al. 2021), and hence formation since 0.7 Gyr after the big bang, poses the most stringent constraints on the masses of black hole seeds. In fact, making the standard assumptions about Eddington-limited accretion, current BH masses in the highest-z quasars appear to rule out the expected $\sim 100\,{\rm M}_\odot$ seeds from Pop III remnants, and instead require more massive seeds ($10^{4-6}\,{\rm M}_\odot$; e.g. Volonteri & Begelman 2010; Volonteri 2012).

As of today, more than 200 quasars have been discovered at redshift $z \geq 6$ (e.g. Fan et al. 2001; Wu et al. 2015; Bañados et al. 2016; Jiang et al. 2016; Matsuoka et al. 2016; Reed et al. 2017; Wang et al. 2017; Yang et al. 2019; Matsuoka et al. 2019b) thanks to the advent of wide-field multiband optical and NIR imaging surveys such as: the *Sloan Digital Sky Survey* (*SDSS*; e.g. Fan et al. 2001), the Canada–France–Hawaii Telescope Legacy Survey (CFHTLS; e.g. Willott et al. 2009), the Panoramic Survey Telescope and Rapid Response System 1 (Pan-STARRS1; e.g. Bañados et al. 2016), the United Kingdom Infrared Telescope Infrared Deep Sky Survey (UKIDSS; e.g. Mortlock et al. 2011), the VISTA Kilo-degree Infrared Galaxy survey (VIKING; e.g. Venemans et al. 2013), the VLT Survey Telescope ATLAS (VST-ATLAS; e.g. Carnall et al. 2015), the Dark Energy Survey (DES; e.g. Reed et al. 2015), the DESI Legacy Imaging Surveys (DELS; e.g. Wang et al. 2017), the UKIRT Hemisphere Survey (UHS; e.g. Wang et al. 2019), and the Hyper Suprime-Cam survey (HSC; e.g. Matsuoka et al. 2016).

At the highest redshifts, there are only eight quasars known at $z \geq 7$ (Mortlock et al. 2011; Bañados et al. 2018; Wang et al. 2018; Yang et al. 2019, 2020b; Matsuoka et al. 2019a, b; Wang et al. 2021) with two of them at $z = 7.5$ (Bañados et al. 2018; Yang et al. 2020a),

★ E-mail: nanni@strw.leidenuniv.nl

and the most distant one at $z = 7.6$ (Wang et al. 2021). This sample of $z \geq 7$ QSOs is still very limited – owing to the opacity of the intervening high-z IGM, distant quasars are brightest redward of their Lyα emission line which is redshifted to NIR wavelengths at $z \geq 7$, making both imaging and spectroscopic observations more challenging. Furthermore, the expected number density of $z \geq 7$ quasars is low (10^{-3} deg^{-2} at $J = 21$; Wang et al. 2019), while the contaminants, mostly Galactic cool dwarfs and early-type galaxies, are far more numerous ($\approx$20 deg^{-2} at $J = 21$). As a result, the success rate in finding quasars in this redshift range is extremely low ($\sim$ 1 per cent based on our experience and private communication with Bañados), and thus requires large amounts of telescope time for spectroscopic confirmation, making it extremely challenging to find more bright $z > 7$ quasars with existing data sets.

On the other hand, the advent of the next generation photometric and spectroscopic telescopes, such as *Euclid* or the *James Webb Space Telescope* (*JWST*), should prove to be a watershed moment in high-redshift quasar studies (Euclid Collaboration 2019). In fact, *Euclid's* wide field IR imaging should enable the discovery of $\sim$100 quasars with $7.0 < z < 7.5$, and $\sim$ 25 beyond the current record of $z = 7.6$, including $\sim$ 8 beyond $z = 8.0$ (Euclid Collaboration 2019), and *JWST* will deliver exquisite spectra of them. Ground-based telescopes will play an essential role in discovering the brighter *Euclid* quasars, whereas fainter $J_{\rm AB} > 21.5$ ones will likely require *JWST*. Although current selection methods based on simple colour-cuts were able to discover most of the $z > 7$ known QSOs (Bañados et al. 2018; Yang et al. 2020a; Wang et al. 2021), their low observational success rate (private communication) is far too low to make confirmation of the on average fainter *Euclid* QSOs feasible, as this would require excessive amounts of ground-based and *JWST* observations. It is thus clear that more efficient selection methods are required.

So far, two different methods for selecting high-z QSOs that are not based solely on colour-cuts have been proposed. The first one is based on Bayesian model comparison (BMC) technique by Mortlock et al. (2012), while the second uses a simpler minimum-χ^2 model fitting method to the quasars' spectral energy distribution (SED; Reed et al. 2017). These methods are based on improved population models for the key contaminants: MLT dwarf types, and compact early-type galaxies, and they both require model colours for each population. The BMC method additionally requires a model for the surface density of each source as a function of apparent magnitude. Although these methods have been successfully used in the past to select high-z QSOs (Mortlock et al. 2011, 2012; Reed et al. 2017), including the VIKING survey (Barnett et al. 2021), they mostly rely on constructing a contaminant model of the entire sky in the colour-range in question to very faint magnitudes, which is a challenging task given our currently poor knowledge about the different types of contaminants. Another quasar search method that has been employed uses the random forests machine learning algorithm in conjunction with colour-cuts for quasar selection and photometric redshift estimation (Schindler et al. 2017, 2018, 2019; Wenzl et al. 2021). While this method has been demonstrated to successfully select quasars at lower-z, its primary drawback is that it cannot properly account for photometric uncertainties.

In this paper, we describe our probabilistic high-z quasar selection technique, which uses density estimation in flux space to compute the probability of being a high-z quasar for each candidate. For density estimation, we use the extreme deconvolution method (XD; Bovy, Hogg & Roweis 2011a, b), which generalizes the familiar machine learning approach of describing a probability density with a mixture of Gaussians to the case of heteroscedastic noise. XD enables one to deconvolve errors for noisy training data to construct the true underlying noiseless probability density, and then reconvolution of the associated noise to evaluate the probability at new arbitrary test locations. In the context of high-z quasar selection, the main merits of this approach are: (1) it is fully Bayesian, and thus similar to the approach of Mortlock et al. (2012) (if the contaminant and quasar models are perfectly known), (2) the contaminant model is fully empirical and requires making no assumptions, (3) it fully accounts for errors in a principled fashion, i.e. noiseless distributions are inferred via deconvolution and then reconvolved with the given target uncertainties. In the end, the target selection/classification problem becomes the task of training good number-density models for both the target population and the contaminant population to maximize the efficiency and completeness of the survey. We applied our target selection technique (hereafter XDHZQSO) to a set of possible high-z candidates that are selected with the use of optical, NIR and MIR surveys, and construct our XDHZQSO quasar targeting catalogue. This catalogue will be used for future spectroscopic follow-up to confirm new high-z QSOs in the NIR ground-based survey area, while this technique provides a better method for classifying and prioritize high-z QSOs candidates in the near future, especially with the advent of *Euclid* in 2022.

This paper is structured as follows. We present the XDHZQSO method in Section 2. In Section 3, we discuss the data used to train our probabilistic classifier, and in Section 4 we describe the construction of the XDHZQSO models from the training data and its application to classify our candidates. In Section 5, we provide a detailed description of the analysis of source completeness and efficiency. In Section 6, we show the results of our code in classifying both the known high-z QSOs in the VIKING survey area, and the known $z > 7$ QSOs on the entire sky. In Section 7, we discuss the limitations of our selection technique, compare it to other methods, and describe various extensions to the basic method described in this paper. We conclude in Section 8. Throughout the paper, we adopt a flat cosmological model with $H_0 = 68.5$ km s^{-1} Mpc^{-1} (Betoule et al. 2014), $\Omega_{\rm M} = 0.3$, and $\Omega_\Lambda = 0.7$. All the magnitudes are given in the AB system, while the uncertainties of our reported measurements are at 1σ confidence level. Throughout the paper, we use F and $\hat{F}$ to indicate the 'true' noiseless and the real noisy fluxes, respectively, whereas we use a generic f to express both the noiseless and the noisy fluxes in the displayed figures.

2 PROBABILISTIC CLASSIFICATION METHOD

High-redshift quasar selection is essentially a probabilistic classification problem in which objects are classified into one of a discrete set of classes, based on their measured physical attributes. These classes can be modelled using a set of objects with class assignments available on which we can train the classification algorithm. Although this is a classical problem in data analysis/machine learning, the physical attributes of astronomical targets are rarely measured without substantial and heteroscedastic measurement uncertainties, and often there is also the problem of sources with no observational coverage in one or more bands of study. Knowing that, classification algorithms for astronomical target selection have to deal with these complications by naturally degrading the probability of an object being in a certain class if the measurement uncertainties imply that the object overlaps several classes.

Consider an object O with 'true' attributes $\{F_i\}$ that we wish to classify into class A or class B. In our specific case, we would like to classify an object O into classes 'high-z QSO' or 'contaminant'

based on its physical noiseless $\{F_i\}$ and noisy $\{\hat{F}_i\}$ attributes (e.g. fluxes, magnitudes, colours, or relative fluxes), and the associated errors $\{\sigma_i\}$. This can be expressed using Bayes' theorem to relate the probability ratio that object O belongs to class A or B to the density in attribute space

$$\frac{P(O \in A|\{\hat{F}_i\})}{P(O \in B|\{\hat{F}_i\})} = \frac{P(O \in A)}{P(O \in B)} \times \frac{p(\{\hat{F}_i\}|O \in A)}{p(\{\hat{F}_i\}|O \in B)}, \quad (1)$$

where the two fractions on the right-hand side are the prior probability ratio and the Bayes factor, respectively. In equation (1), we distinguish between discrete probabilities P and continuous probabilities p. The $p(\{\hat{F}_i\}|O \in A)$ factor in the numerator of the right-hand side of equation (1) is the density in attribute space evaluated at the targets's attributes $\{\hat{F}_i\}$, while $P(O \in A)$ is proportional to the total number of A objects in a prior probability. The denominator $p(\{\hat{F}_i\})$ is a normalization factor, and expresses the total probability that the object O belongs to either class A or class B. It is easy to see that this probability is a true probability since it always lies between zero and one, and the sum of the probabilities for the two classes is equal to one.

Measurement uncertainties are handled in this framework through marginalization over the 'true' properties $\{F_i\}$ given the observed ones $\{\hat{F}_i\}$ and the measurement-uncertainty distribution $p(\{\hat{F}_i\}|\{F_i\})$:

$$p(\{\hat{F}_i\}|O \in A) = \int \mathrm{d}\{F_i\} p(\{F_i\}|O \in A) p(\{\hat{F}_i\}|\{F_i\}). \quad (2)$$

We take $p(\{\hat{F}_i\}|\{F_i\})$ to be Gaussian, which is an extremely good approximation for flux measurements. XD provides a simple mechanism to (1) infer the true underlying 'noise deconvolved distribution' $P(O \in A|\{\hat{F}_i\})$, as well as (2) performs the convolution integral in equation (2). Since the model is a mixture of Gaussians and the errors are Gaussian, the normally complex operations of deconvolution/convolution reduce to trivial algebraic operations.

Compared to other probabilistic selection methods, the great advantage of our approach is that the poorly understood contaminants are modeled fully from the data,[1] rather than relying on empirical models (e.g. Mortlock et al. 2012; Barnett et al. 2021), and the contaminant classes are all grouped into a single all-inclusive contaminant class. In this way, the density models for the contaminant class can be simply trained using real data from the entire sky. This method was already applied in the past to select *SDSS* QSOs (Bovy et al. 2011b; Bovy et al. 2012), and was shown to be effective even in the challenging redshift range $2.5 \leq z \leq 3$ where the stellar contamination is significant.

3 TRAINING DATA

To construct probability density models we trained on either real or simulated photometry, depending on whether we are considering 'contaminants' or 'quasars'. Contaminants were trained on 1076 deg^2 of overlapping imaging from VIKING ($YJHK_s$), DECaLS (grz), and *unWISE* ($W1W2$).[2] In Table 1, we summarize the properties of the three surveys we used for our selection. The quasar models were trained on synthetic photometry from the McGreer et al. (2013) 'simqso' simulator.[3] This section describes the data used to train these density classification models.

Table 1. Survey properties.

Survey	Filters	5σ depth
VIKING	$ZYJHK_s$	23.1, 22.3, 22.1, 21.5, 21.2
DECaLS	grz	23.95, 23.54, 22.50
un*WISE*	$W1W2$	20.72, 19.97

3.1 Contaminant data

The contaminant training set is generated using photometry from deep optical, and near- and mid-IR imaging surveys.

At NIR wavelengths, we used Y, J, H, and K_s bands coming from VIKING DR4. The VIKING data were obtained from the VISTA Science Archive.[4] For optical bands, we mainly used data from the DESI Legacy Imaging Surveys (DELS),[5] which combines three different imaging surveys: the DECaLS, the Beijing-Arizona Sky Survey (BASS; e.g. Zou et al. 2019), and the Mayall z-band Legacy Survey (MzLS). These three surveys jointly image ∼14 000 deg^2 of the extragalactic sky visible from the Northern hemisphere in three optical bands (g, r, and z). The sky coverage is approximately bounded by $-18^\circ < \delta < +84^\circ$ in celestial coordinates, and $|b| > 18^\circ$ in Galactic coordinates, and it overlaps with most (≈ 80 per cent) of the VIKING survey footprint. An overview of the DELS surveys can be found in Dey et al. (2019). When available, we also included Pan-STARRS (PS1) photometric data in our selection, which provides 3π sky coverage (≈ 70 per cent overlap with the VIKING footprint) in five different filters: g_{PS1}, r_{PS1}, i_{PS1}, z_{PS1}, and y_{PS1}. As described below, these data were used to further refine our training catalog. In the MIR, we used the $W1$ and $W2$ bands coming from the un*WISE* release (Schlafly, Meisner & Green 2019), that comes from the coaddition of all publicly available 3–5 μm *WISE* imaging (Wright et al. 2010), including that from the ongoing NEOWISE (Mainzer et al. 2011) post-cryogenic phase mission. The steps used to construct our catalogue are illustrated schematically in Fig. 1, which we describe in detail in the following.

As we are interested in finding $6 \leq z \leq 8$ QSOs, we used the J band as the 'detection band' to construct our contaminant training sample. In fact, at the very high-redshift ($z > 7$) the Ly α drop falls in the Y band, preventing the detection of very high-z QSOs, while the VIKING J band reaches a depth of 22.1 (at 5σ). So, we selected all the sources with J band signal-to-noise ratio SNR (J) ≥ 5. We also removed bright sources ($J < 17$), as we found they were often artifacts or bright stars, after performing a visual inspection of them. Then, we cross-matched the VIKING catalogue with the DELS, PS1, and unWISE ones, using a radius 2 arcsec. For sources covered by the DELS footprint but with no counterpart detected in the survey within 2 arcsec, we performed forced photometry on the DECaLS images with an aperture radius 1.5 arcsec. At this stage, since $z \geq 6$ QSOs drop out in the bluest optical filters, we further required our objects to have SNR(g, r) < 3,[6] and, when available, SNR(g_{PS1},

[1] However, the high-z QSOs are trained on empirical models.

[2] To compute the area covered by the sources in our sample we used the *healpy* PYTHON package, based on the Hierarchical Equal Area isoLatitude Pixelization (HEALPIX). We used *healpy* to subdivide a spherical surface in 200 pixels, in which each pixel covers the same surface area as every other pixel, and summed the areas of the pixels that includes one or more sources from the VIKING survey area.

[3] https://github.com/imcgreer/simqso/

[4] http://horus.roe.ac.uk/vsa/

[5] https://www.legacysurvey.org/

[6] Sources detected in DELS have already forced photometry for the DECaLS-grz and the un*WISE*-$W1W2$ filters.

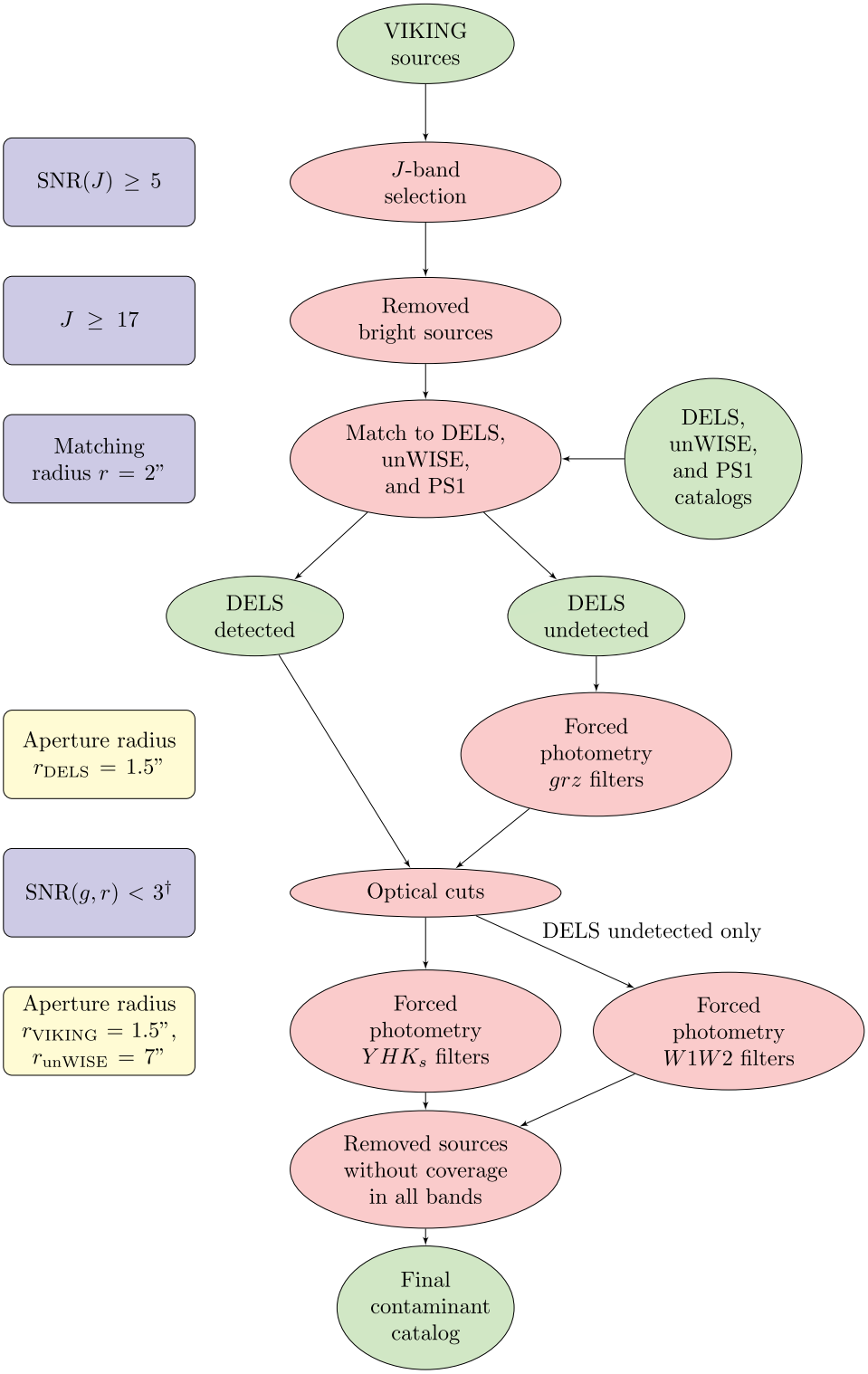

Figure 1. General steps (red ellipses) performed to construct the contaminant training sample. The blue boxes represent the conditions that the sources must satisfy to make it to the next step, while yellow boxes provide more information about some specific steps. After the match with other surveys (DELS, unWISE, and PS1), sources are divided into two sub-catalogues depending on their DELS counterpart: sources with a DELS detected counterpart (DELS detected), and sources with no detected counterpart but with DELS coverage (DELS undetected). Sources with neither DELS counterpart nor DELS coverage are simply removed. † At this step we also removed sources with SNR($g_{\rm PS1}$, $r_{\rm PS1}$) $\geq$ 3, or SNR($i_{\rm PS1}$) $\geq$ 5 and $i - z < 2$, when these data are available.

Table 2. Selection criterion on the 'contaminant' training catalogue.

Data sample	Number of sources
VIKING catalogue	94 819 861
SNR(J) $\geq$ 5	45 968 999
$J \geq 17$	44 191 759
VIKING cross-matched[a]	36 057 930
SNR(g, r) < 3[b]	2871 420
Sources with data in all bands	1902 071

[a]Specifically, there are 33 633 899 sources with a DELS detected counterpart, and 2424 031 sources with no DELS detected counterpart but covered by the DELS survey.
[b]At this step we also removed all the sources with SNR($g_{\rm PS1}$, $r_{\rm PS1}$) ≥ 3, or SNR($i_{\rm PS1}$) ≥ 5 and $i - z < 2$, when these data are available.

$r_{\rm PS1}$) < 3. We also removed objects with SNR($i_{\rm PS1}$) ≥ 5 and $i - z < 2$, when these data were available. For the surviving sources, we performed forced photometry on the VIKING images (*YHK*$_s$ filters), using an aperture radius 1.5 arcsec, while we also performed forced photometry on the un*WISE* images, with an aperture radius 7 arcsec, for those sources with no DELS detected counterpart. Finally, we removed sources that have no coverage in all the requested filters (VIKING-*YHK*$_S$, DECaLS-z, and un*WISE*-*W*1*W*2).[7]

The resulting final 'contaminant' training catalogue contains 1902 071 sources, while the number of sources that survived each filtering step are presented in Table 2. Among the final sources, we identified eight known $6 \leq z \leq 7$ QSOs, indicating that the contamination of the contaminant training set with high-z quasars is small. Therefore, we did not remove these known QSOs from the training set.

3.2 Quasar data

We used a sample of 440 000 $6 \leq z \leq 8$ QSOs simulated from the 'simqso' code from McGreer et al. (2013), using the updated version described in Yang et al. (2016). The simqso code was used to generate a grid with a uniform distribution in redshift over the range $6 \leq z \leq 8$, and in magnitude over the range $17 \leq J \leq 22.5$. Assuming that the QSO spectral energy distributions (SEDs) do not evolve with redshift (Kuhn et al. 2001; Yip et al. 2004; Jiang et al. 2006; Bañados et al. 2018), the quasar spectrum is modelled as a power-law continuum with a break at 1200 Å. For redder wavelength coverage, we added four breaks at 2850, 3645, 6800, and 30 000 Å. The slope (α_λ) from 1200 to 2850 Å follows a Gaussian distribution with mean $\mu(\alpha_{1200}) = -0.5$ and dispersion $\sigma(\alpha_{1200}) = 0.3$; the range from 2850 to 3645 Å has a slope drawn from a Gaussian distribution with $\mu(\alpha_{2850}) = -0.6$ and $\sigma(\alpha_{2850}) = 0.3$; from 3645 to 6800 Å we adopted a Gaussian with $\mu(\alpha_{3645}) = 0.0$ and $\sigma(\alpha_{3645}) = 0.3$; finally, from 6800 to 30000 Å, we used $\mu(\alpha_{6800}) = 0.3$ and $\sigma(\alpha_{6800}) = 0.3$. These different break points and power-law exponents are designed to reproduce the template from Selsing et al. (2016). The parameters of emission lines are derived from the composite quasar spectrum from (Glikman, Helfand & White 2006), and the lines are added to the continuum as Gaussian profiles, where the Gaussian parameters (wavelength, equivalent width, and full with half-maximum) are drawn from Gaussian distributions. These distributions recover trends in the mean and scatter of the line parameters as a function of continuum luminosity, e.g. the Baldwin effect (Baldwin 1977), and blueshifted lines (Gaskell 1982; Richards et al. 2011). The simulator also models absorption from from neutral hydrogen absorption in Ly α forests based on the work of Worseck & Prochaska (2011). As a reference, we provide in Fig. 2 the mean spectrum of 20 000 $z \sim 6$ simulated QSOs (red line), and the spectra corresponding to the 16th and 84th percentiles (blue lines), normalized at 1450 Å. The final noiseless photometry of simulated QSOs is derived from the model spectra by integrating them against the respective filter curves.

[7]Although, XD can menage the problem of sources with missing data by using a very large uncertainty variance for them, we decided to train our models using the best data available (i.e. removing sources with no coverage in all the filters of study). We plan to use the XD feature that allows to deal with missing data in future works.

4 XDHZQSO DENSITY MODEL

To estimate the density of contaminants and quasars in flux space [the $p(\{\hat{F}_i\}|O \in A)$ factor from equation (1)], we used the XDGMM[8] implementation of extreme deconvolution from Holoien, Marshall & Wechsler (2017). XDGMM is a PYTHON package that utilizes the scikit-learn API (Pedregosa et al. 2011; Buitinck et al. 2013) for Gaussian mixture modelling. It performs density estimation of noisy, heterogenous, and incomplete data and uses the XD algorithm[9] (Bovy et al. 2011b) for fitting, sampling, and determining the probability density at new locations. As described by Bovy et al. (2011b), XD models the underlying, deconvolved, distribution as a sum of N Gaussian distributions, where N is a model complexity parameter that needs to be set using an external objective. It assumes that the flux uncertainties are known, as is in our case, and consists of a fast and robust algorithm to estimate the best-fitting parameters of the Gaussian mixture. In Section 4.2, we follow the approach used by Bovy et al. (2011b) to construct the flux density model of the two classes.

Finally, since Gaussian mixture models are unit-normalized, to compute the probability of an object belonging to a certain class, we require a separate prior to get the correct relative weighting of the two populations. In practice, we need to estimate the number counts of both quasars and contaminants [the $P(O \in A)$ factor from equation (1)]: i.e. these are the prior factors of our Bayesian approach. For the contaminants, we compute this factor empirically from the number counts (J-band magnitude distribution of contaminants), while for the quasars we derived them from the high-z QSO luminosity function. However, to derive the true number counts for the QSOs, which includes the survey incompleteness at the faint end, we used the empirical data to compute the incompleteness for the VIKING survey, and apply it to the QSO number counts. In Section 4.3, we provide details about the computation of these prior factors.

4.1 The binning approach

The full model consists of fitting the probability density [the $p(\{\hat{F}_i\}|O \in A)$ and $p(\{\hat{F}_i\}|O \in B)$ factors from equation (1)] in a number of bins in J-band magnitude for the two classes of objects. We opted to bin in J band because the probability density of quasars will have a dominant power-law shape corresponding to the number counts as a function of apparent magnitude, whereas the colour distribution is much flatter. While the latter can be represented well by mixtures of Gaussian distributions, the power-law behaviour cannot without using large numbers of Gaussians. Thus the slow variation

[8]https://github.com/tholoien/XDGMM
[9]https://github.com/jobovy/extreme-deconvolution

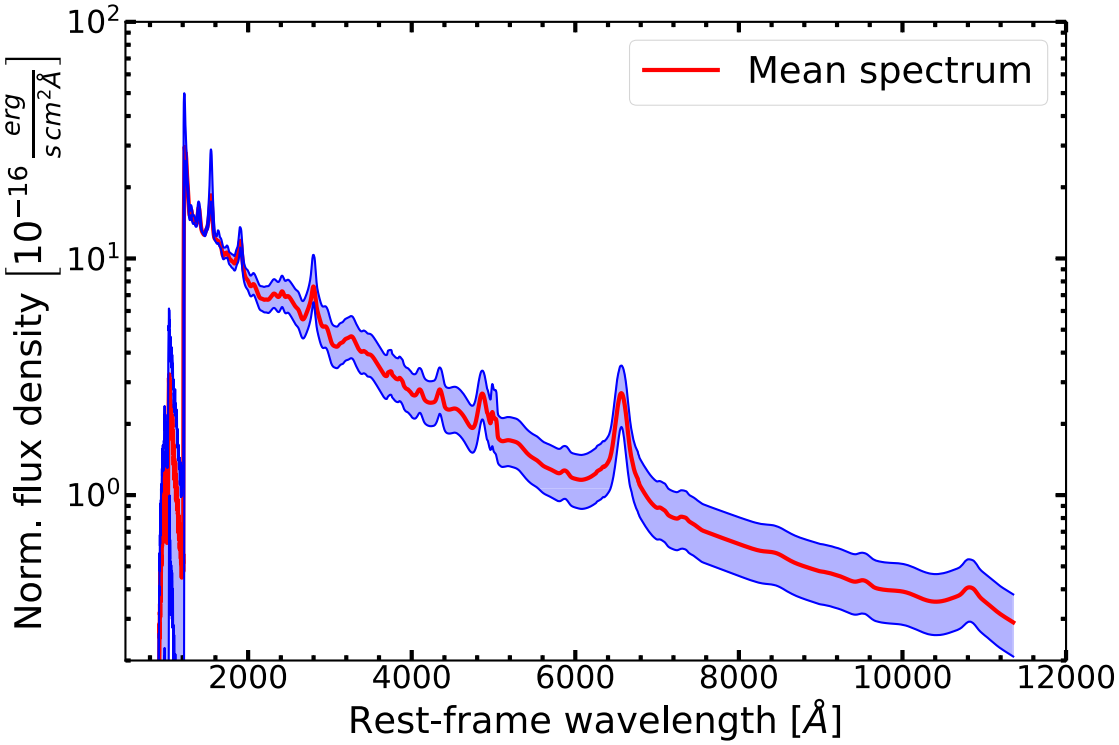

Figure 2. Rest-frame mean spectrum of 20 000 $z \sim 6$ simulated QSOs (red line), and the spectra corresponding to the 16th and 84th percentiles (blue lines), normalized at 1450 Å. The spectra are modelled as a power-law continuum with a break at 1200 Å, so to reproduce the template from Selsing et al. (2016), while the parameters of emission lines are derived from the composite quasar spectrum from (Glikman et al. 2006), and the lines are added to the continuum as Gaussian profiles.

of the colour distributions with magnitude is captured by our model, since we use narrow bins in J-band magnitude.

The full contaminant model consists of 50 overlapping bins where the right edges are uniformly distributed in the range $J = 20 - 22.5$ with a step of 0.05 mag, while the width is given by a broken sigmoid function:

$$w = bw + (bs_1 - bw)\frac{1}{1 + e^{\frac{J_{\rm bre} - m_{\rm th_1}}{\Delta m}}} \quad \text{for} \quad J_{\rm bre} \leq 22,$$
$$w = bw + (bs_2 - bw)\frac{1}{1 + e^{\frac{J_{\rm bre} - m_{\rm th_2}}{\Delta m}}} \quad \text{for} \quad J_{\rm bre} > 22 \qquad (3)$$

where, $J_{\rm bre}$ is the J-band bin right edge, $bw = 0.1$ represents the minimum bin width and $bs_1 = 5$, $bs_2 = 1$ represent the maximum bin widths in the two J-band ranges, $m_{th_1} = 21$, $m_{th_2} = 22$, and $\Delta_m = 0.1$. The broken sigmoid for the contaminants is shown in Fig. 3. The use of a variable bin width is driven by the need of having a model that is as continuous as possible, as the XD fits can jump between local maximums. In fact, this procedure guarantees that many (> 20 per cent) of the objects in the bins overlap for adjacent bins, and thus the model varies smoothly. Furthermore, the use of a broken sigmoid guarantees that both at the bright and faint ends, where fewer objects are present, the bins are large enough to contain a sufficient number of sources. In fact, we have >2000 training objects in each bin to build the contaminant models.

As for the quasar model, we used 11 uniform spaced bins with a width 0.5 mag in the range $J = 17 - 22.5$, and we further divided the quasar class into three subclasses corresponding to 'low-redshift' ($6 \leq z \leq 6.5$), 'medium-redshift' ($6.5 \leq z \leq 7$), and 'high-redshift' ($7 \leq z \leq 8$) quasars, constructing a QSO model for each bin. We opted

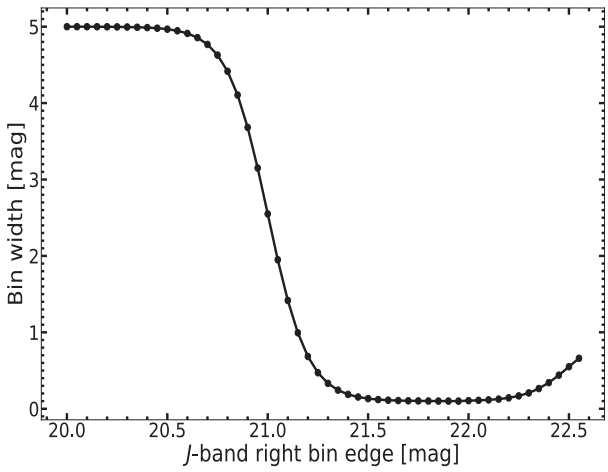

Figure 3. Double sigmoid function that displays the right edges and the width of the bins used to train the contaminant model.

to divide the QSO into these three redshift bins, instead of working with a broad $6 \leq z \leq 8$ bin, for the following reasons:

(i) As shown in Section 5.1, the efficiency and completeness of our selection method strongly depends on the z-bin in question owing to the changing overlap between quasars and contaminants.

(ii) While the $6 \leq z \leq 7$ range has been largely investigated in the past, few objects have been found at $7 \leq z \leq 8$, making it the highest priority range that we are interested in investigating.

(iii) Spectroscopic wavelength coverage is different for different instruments, with the dividing line between optical and near-IR

spectrographs typically occurring around ≈10 000 Å ($z_{\mathrm{Ly}\,\alpha} = 7.2$). Thus, not all the $6 \leq z \leq 8$ QSOs can simply be confirmed with a single instrument, and multiple instruments could be required to confirm candidates over such a broad redshift range. Hence, the redshift bins we adopted also facilitates in efficiently conducting follow-up observations.

However, in the future we plan to introduce the redshift as one of the modelled quantities as done by Bovy et al. (2012), so that one would no longer needs to construct models in different redshift bins, as this approach also provides photometric redshifts, which can be used to select candidates over any desired redshift range.

4.2 Construction of the model

The XD code fits for all the *J*-band magnitude bins for a given class are initialized using the best-fitting parameters for the previous bin, so to guarantee the continuity of the mode. The starting bin (the one that is not initialized) is the closest to $J = 21$, where we know we always have a quite large sample of objects ($>10^5$) for the training. Hereafter, we describe the model in a single bin first for a single example class, using the contaminant class as the example.

In a single bin in *J*-band magnitude, we separate the absolute flux from the flux relative to the *J* band in the likelihood in equation (1) as follows:

$$p(\{\hat{F}_i\}|O \in \text{“cont.''}) = p(\{\hat{F}_i/\hat{F}_J\}|\hat{F}_J, O \in \text{“cont.''}) \times p(\hat{F}_J|O \in \text{“cont.''}), \quad (4)$$

where $\{\hat{F}_i\}$ are the *z*, *Y*, *H*, *K*, *W*1, *W*2 fluxes, $\{\hat{F}_i/\hat{F}_J\}$ are the fluxes relative to *J* band, and $\hat{F}_J$ is the *J*-band flux. We model the two factors of the right-hand side of equation (4) separately.

We modelled the $p(\{\hat{F}_i/\hat{F}_J\}|\hat{F}_J, O \in \text{“cont.''})$ factor using XD in narrow bins in *J*-band magnitude. We use relative fluxes rather than colours since the observational uncertainties are closer to Gaussian for relative fluxes than they are for colours. Also, for sources where the flux measurement can be negative the magnitudes are badly behaved, while relative fluxes remain well behaved in this case. To evaluate the XD probabilities during training, we always convolved the underlying model with the object's relative-flux uncertainties assuming that they are Gaussian distributed, such that the convolution of the Gaussian mixture with the Gaussian uncertainty results in another Gaussian mixture. Although the ratio of two noisy Gaussian deviates is not itself Gaussian distributed, Gaussianity is a good approximation provided that the *J*-band flux errors are small. The validity of this approximation is discussed further in Appendix A. Note also that since all other fluxes are divided by the *J*-band flux, the resulting uncertainties are covariant, and we provide the functional form of this covariance matrix in A. To train for the QSO models, since the simulated quasar fluxes are noiseless, we simply need to fit their flux densities without deconvolving to derive the underlying deconvolved quasar model. However, to avoid singular inverse variances for the effectively noiseless model data, we added a tiny error (0.01) to the simulated noiseless relative fluxes drawn from a Gaussian distribution, and used for consistency this small value of the error as the input error on the photometry in the XD code. In Fig. 4, we show the relative-flux relative-flux diagrams of our training data: the contaminants are displayed using black contours, while a subsample (5000) of simulated $6 \leq z \leq 8$ QSOs are shown as coloured points. For display purposes, we added to the displayed quasars real errors drawn from a noise model based on our contaminant catalogue which is described in Appendix B.

We modelled the 6D deconvolved relative fluxes $\{F_i/F_J\}$, using 20 Gaussian components. The number 20 was chosen after performing XD fits with 10, 15, 20, and 25 components. While fits with less than 20 components overly smoothed the observed distribution, models with more than 20 components used the extra components to fit extremely low significance features in the observed distribution. The same number of components was also adopted by Bovy et al. (2011b). Similarly, we also used 20 Gaussian components to fit for the quasar models.

To provide a visual example of the model generated by the XDHZQSO code, we display in Fig. 5 the $20.67 < J < 21.2$ deconvolved contaminant model (black contours) compared to the $20.5 < J < 21.0$ QSO models in the three redshift bins: $6 \leq z \leq 6.5$ (blue), $6.5 \leq z \leq 7$ (green), and $7 \leq z \leq 8$ (red). To generate the displayed samples, we drew 50 000 sources from the deconvolved contaminant model, and 50 000 objects from the three redshift-bins deconvolved QSO models. It is apparent that the large overlap between the contaminant and the $6.5 \leq z \leq 7$ and $7 \leq z \leq 8$ QSO contours will greatly lower the efficiency in selecting QSO candidates in these two redshift ranges, as better explained in Section 5.1. To asses the quality of our contaminant deconvolved models, we sampled the deconvolved models in each *J*-band bin,[10] re-added the errors to the deconvolved fluxes following our noise modelling procedure described in Appendix B, and compared the relative-flux distribution of the reconvolved sample with the original real noisy data. In Fig. 6, we compare a simulated set of samples (red contours) from the deconvolved $20.67 < J < 21.2$ contaminant model with the real data distribution (black), while in Fig. 7 we compare the same simulated sample after adding the errors, following Appendix B (red), with the real contaminant distribution (black). It is apparent that, after re-adding the errors to the deconvolved quantities, we obtain a distribution that is consistent with the $20.67 < J < 21.2$ real data.

4.3 Computation of the priors

The second factor of equation (4), $p(\hat{F}_J|O \in \text{“cont.''})$ is expressed as

$$p(\hat{F}_J|O \in \text{“cont.''}) = \frac{NC_{\text{cont.}}(\hat{F}_J)}{(NC_{\text{cont.}}(\hat{F}_J) + NC_{\text{Q}}(\hat{F}_J))}, \quad (5)$$

where $NC_{\text{cont.}}(\hat{F}_J)$ and $NC_{\text{Q}}(\hat{F}_J)$ are the number counts of contaminants and QSOs at a specific $\hat{F}_J$, respectively. Since the denominator $(NC_{\text{cont.}}(\hat{F}_J) + NC_{\text{Q}}(\hat{F}_J))$ factors out in equation (1), $p(\hat{F}_J|O \in \text{“cont.''})$ can be expressed simply by the number counts of contaminants (or quasars) as a function of apparent magnitude, and is always expressed in units of deg^{-2}. For the contaminant class, we modelled the number counts directly using the number counts of the training data, by fitting the histogram of *J*-band magnitude number counts per square degree. We used a 40-order polynomial to perform a robust fit to the range $J \leq 21.4$, while at $J > 21.4$ we used a cubic spline to interpolate the histogram, namely to capture the drop-off due to catalogue incompleteness. In order to model the effect of the incompleteness on the real data distribution, we fit a power law to the range $20.7 \leq J \leq 21.4$:

$$f(J) = cJ^{\alpha} \quad \text{for} \quad 20.7 \leq J \leq 21.4, \quad (6)$$

where $log(c) = -95.3$ and $\alpha = 73.0$, and extrapolated this power-law fit to $J > 21.4$. The ratio between the value given by the power

[10]For each bin we sampled a number of sources equal to the number of real VIKING sources from that bin.

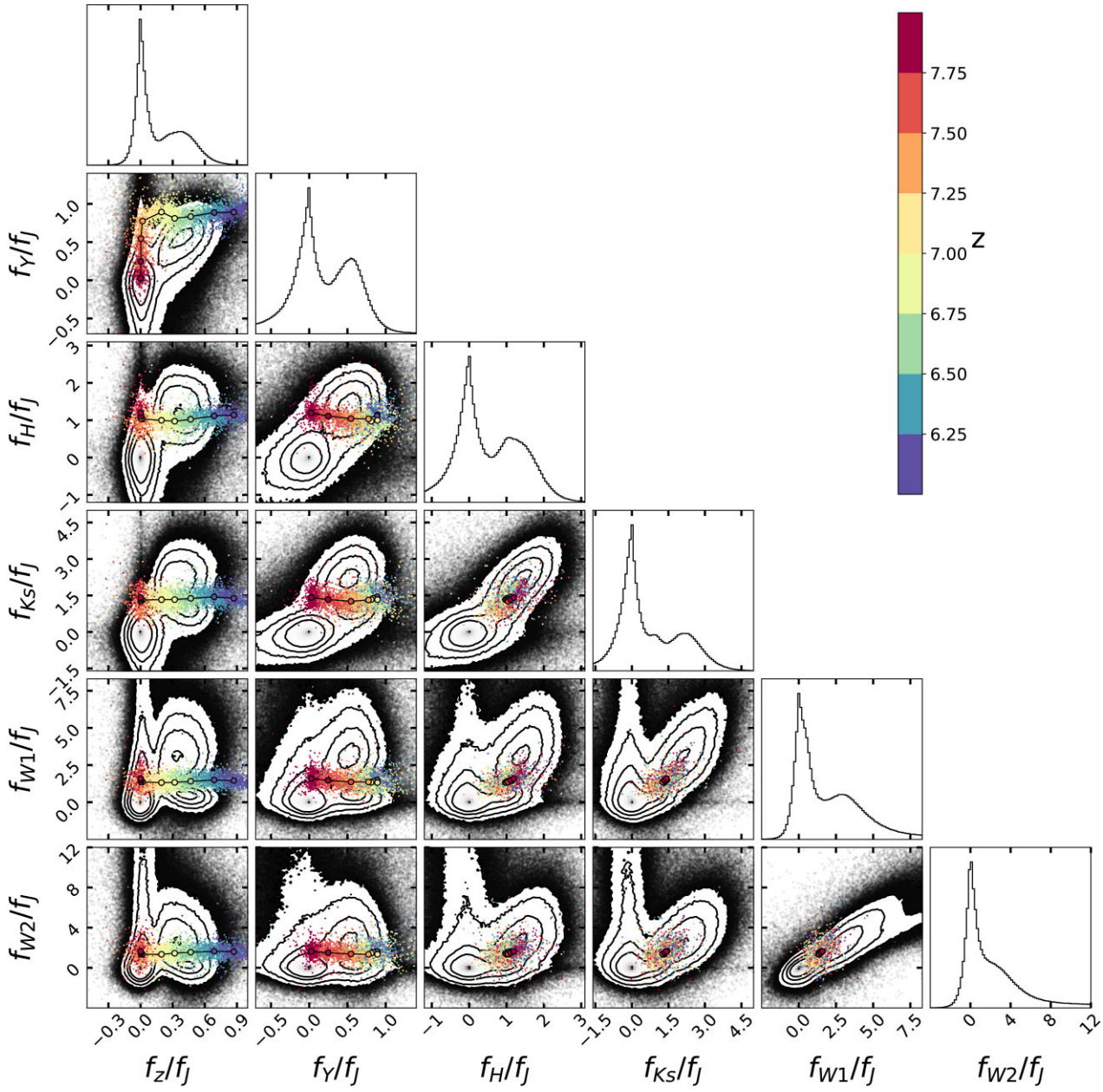

Figure 4. Noisy relative-flux plots for both the contaminant (black contours and points) and a sub-sample (5000) of high-z QSO training data (coloured points). The colourbar shows the redshift of the simulated QSOs, while the labelled quantities are relative fluxes (i.e. fluxes in different bands divided by the J-band flux). For display purposes, we added to the simulated noiseless quasars the real errors coming from our contaminant catalogue as explained in Appendix B, while the black line and coloured filled circles represent the colour–redshift relation predicted using our simulated QSOs. Although we do not know the real nature of our contaminants, we expect that most of them are cool brown dwarves and early type galaxies.

law and the cubic spline interpolated number counts gives us the incompleteness correction term to apply to our QSO number counts at $J > 21.4$. We show in Fig. 8 (black line) the $p(\hat{F}_J|O \in \text{“}cont.\text{”})$ factor.

For the ‘quasar’ class, we used a model for the $z \sim 6.7$ quasar luminosity function (LF) from Wang et al. (2019) to compute the number density of quasars as a function of the apparent J-band magnitude, in the three redshift bins ($6 \leq z \leq 6.5$, $6.5 \leq z \leq 7$, and $7 \leq z \leq 8$). This LF is characterized by a double power law:

$$\Phi(M_{1450}, z) = \frac{\Phi^*(z)}{10^{0.4(\alpha+1)(M_{1450}-M^*)} + 10^{0.4(\beta+1)(M_{1450}-M^*)}}, \quad (7)$$

where M_{1450} is the absolute magnitude at 1450 Å, α and β are the faint-end and bright-end slopes, respectively, M^* is the characteristic magnitude, and $\Phi^*(z) = \Phi^*(z=6) \times 10^{k(z-6)}$ is the normalization, where $k = -0.72$ as measured by Jiang et al. (2016) for $5 < z < 6$ QSOs. We fixed the four parameters to the $z \sim 6.7$ LF measured by Wang et al. (2019): $\alpha = -1.9$, $\beta = -2.54$, $M^* = -25.2$, and $\log_{10}(\Phi^*) = -8.5$. To express the LF as a function of J-band apparent magnitude we convert the M_{1450} to J-band magnitude (m_J) using:

$$m_J = M_{1450} + DM + k\text{−corr}, \quad (8)$$

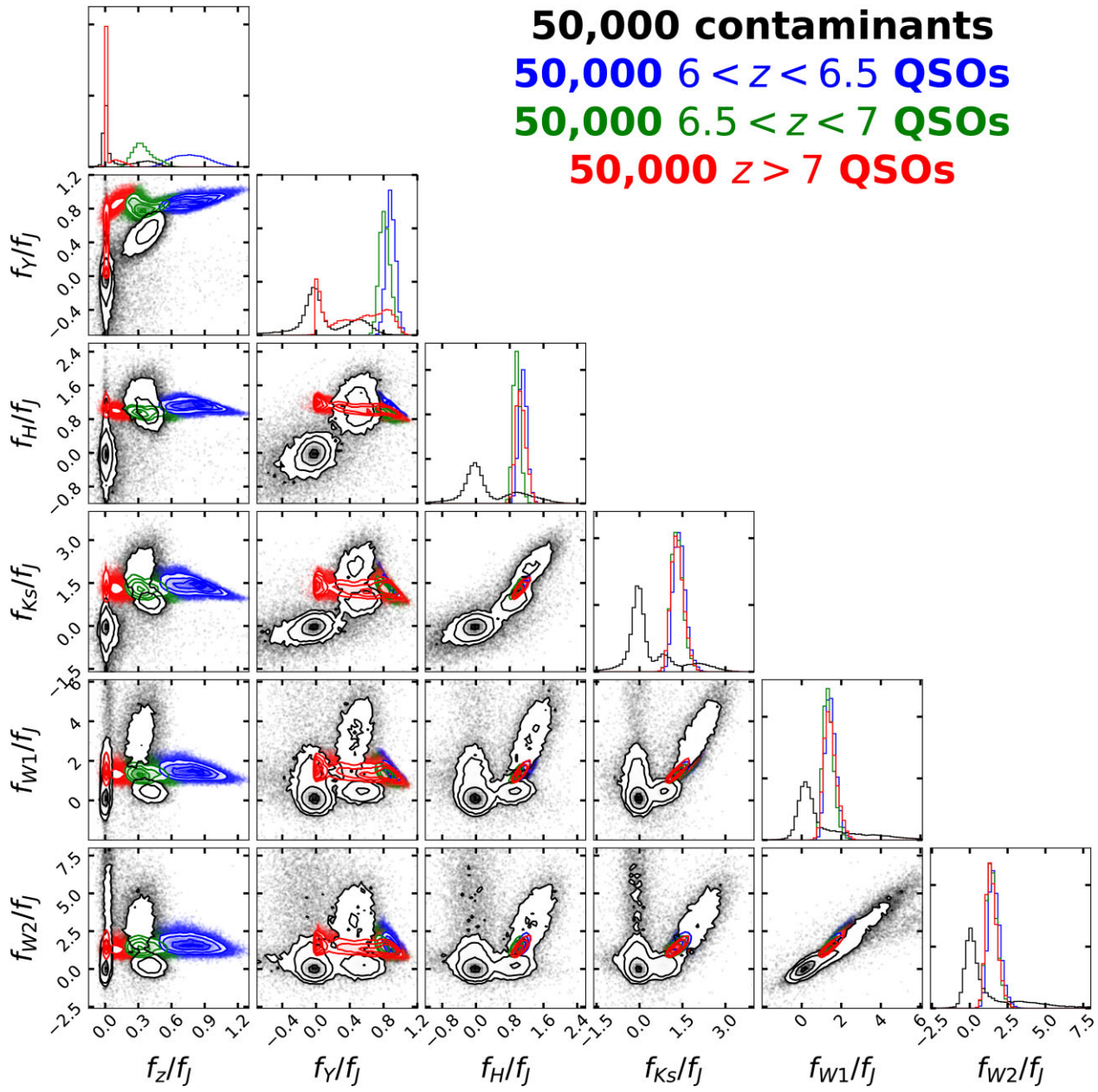

Figure 5. Noiseless relative-flux relative-flux contours for the $J = 20.67 - 21.2$ deconvolved contaminant model (black), and for the deconvolved $J = 20.5 - 21.0$ $6 \leq z \leq 6.5$ (blue), $6.5 \leq z \leq 7$ (green), and $7 \leq z \leq 8$ (red) QSO models. The labelled quantities are relative fluxes (i.e. fluxes in different bands divided by the *J*-band flux). To generate the displayed samples, we sampled 50 000 sources from the contaminant model, and 50 000 objects from each of the three QSO models.

where *DM* is the distance module and *k*-corr the *k*-correction from Richards et al. (2006):

$$k-\text{corr} = -2.5(1+\alpha_\nu)\log_{10}(1+z) - 2.5 * \alpha_\nu \log_{10}\left(\frac{145(\text{nm})}{1254(\text{nm})}\right), \quad (9)$$

where $\alpha_\nu = -0.5$. Then, we multiplied in the survey incompleteness at $J > 21.4$ that we computed from the contaminant distribution. We show in Fig. 8, the $p(\hat{F}_J|O \in$ "$6 \leq z \leq 6.5$ quasar$''$) factor (blue line), the $p(\hat{F}_J|O \in$ "$6.5 \leq z \leq 7$ quasar$''$) factor (green line), and the $p(\hat{F}_J|O \in$ "$7 \leq z \leq 8$ quasar$''$) factor (red line).

5 HIGH-z QSO SELECTION

In this section, we present the XDHZQSO source classification for all the objects selected by our initial cuts described in Section 3.1. This catalogue was also used to train the contaminant model as described in Section 4, since we argued that the fraction of high-z QSOs contained in this catalog is negligible. Using the models of quasar and contaminant deconvolved relative fluxes, we computed the probability that every object is a high-z QSO or a contaminant using equation (4). Specifically, we used the models from the previous section as follows. For an object with *J*-band magnitude *J*, we first found the bin whose midpoint is the closest to this magnitude. Then, we used this bin

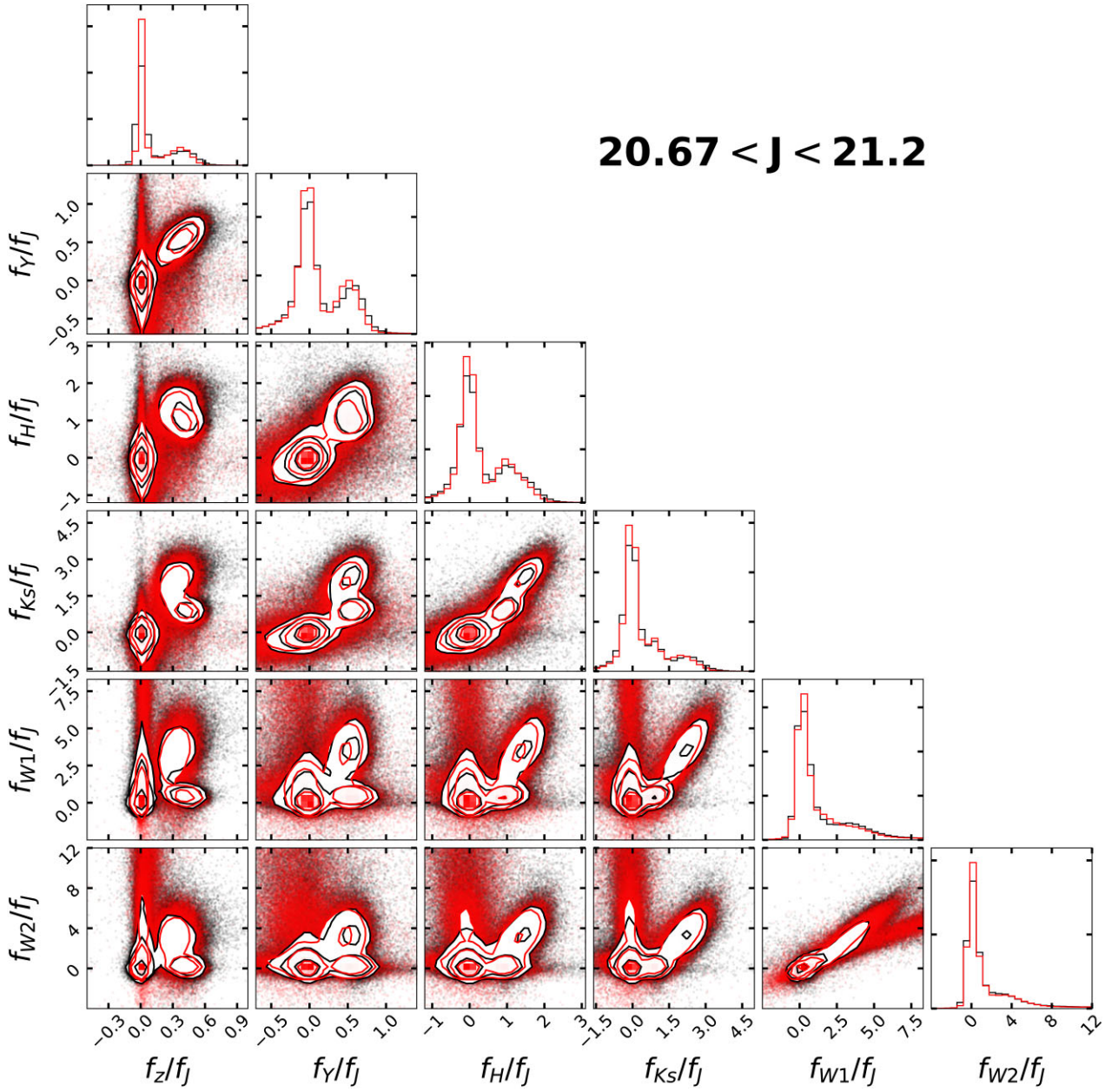

Figure 6. Relative-flux relative-flux contours for the (noiseless) deconvolved $20.67 < J < 21.2$ contaminant model (red), compared to the real (noisy) data distribution (black). The labelled quantities are relative fluxes (i.e. fluxes in different bands divided by the *J*-band flux). Overall, the red contours are tighter compared to the black ones showing the efficacy of XDHZQSO in deconvolving the noisy distributions.

to evaluate the relative-flux density $p(\{\hat{F}_x/\hat{F}_J\}|\hat{F}_J, O \in \text{“}cont.\text{”})$ for this object's relative fluxes[11] by convolving the underlying 20 Gaussian mixture model with the object's uncertainties. This uncertainty convolution is simply adding the object's uncertainty covariance to the intrinsic model covariance for each component.

Finally, we evaluated the number density as a function of the object's apparent magnitude in *J* band, using the interpolated relations described in Section 4.2. We did this for each of the classes (contaminant and the three quasar classes) and compute the probabilities using equation (1).

[11]Where $\hat{F}_x$ is the flux in an arbitrary band other than *J*.

In Fig. 9, we show the distribution of XDHZQSO quasar probabilities for the sources we classified in the VIKING survey area in the three redshift bins defined in Section 4.1. Since the catalogue is expected to contain mostly contaminant sources, the probability distribution is peaked at zero in each redshift bin, with a few exceptions at higher probabilities that represent our best candidate quasars for future spectroscopic confirmation. It is also apparent that the number of the best candidates for spectroscopic follow-up (i.e. those with $P_{\rm QSO} > 0.1$) decreases as the redshift increases. This results from the combination of two factors: (1) the number density of QSOs decreases as redshift increases, (2) the overlap in the relative-flux-relative-flux space between the higher-z QSOs and the contaminants is larger, in particular in the $6.5 \leq z \leq 7$ range (see Fig. 5).

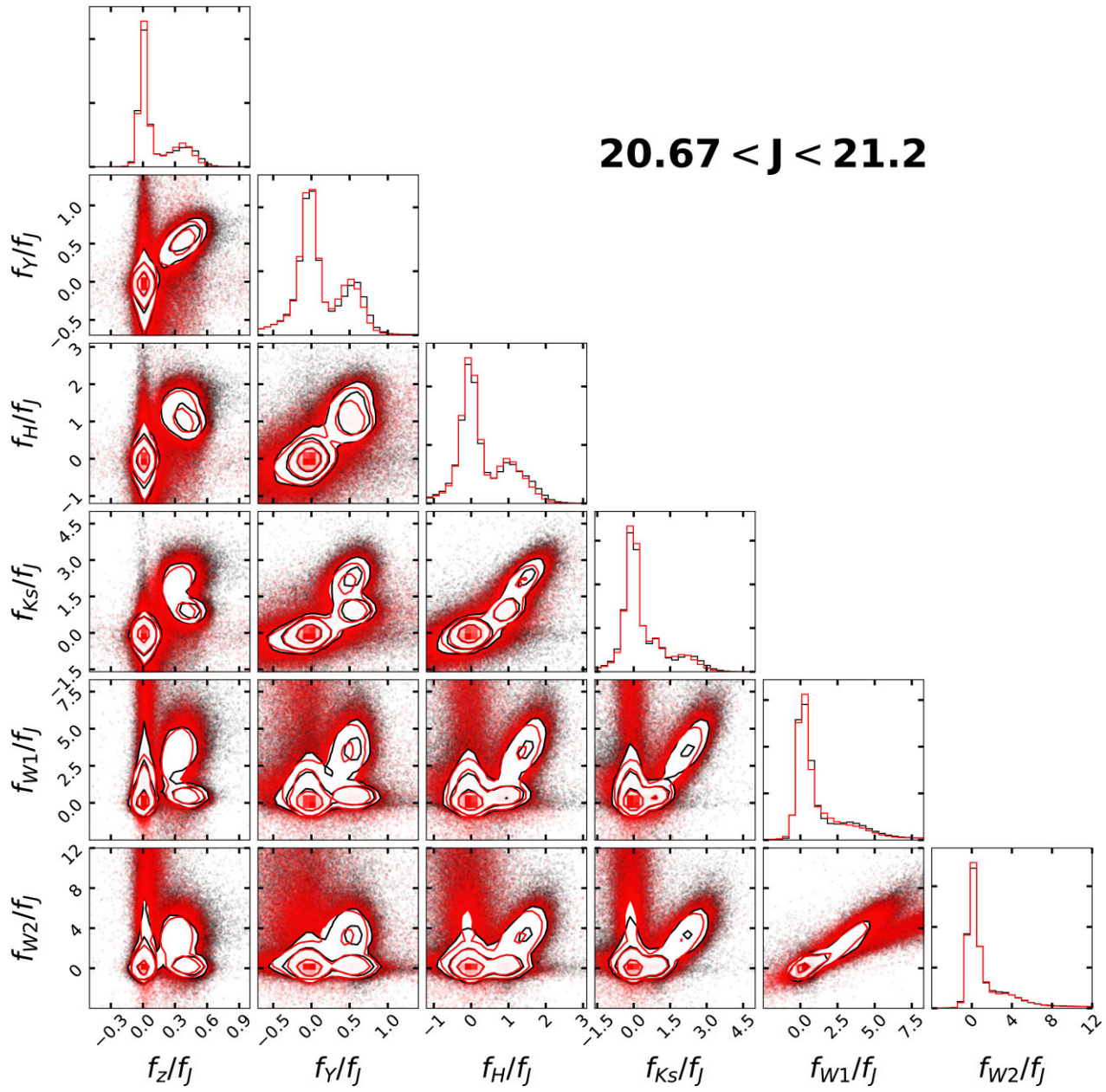

Figure 7. Relative-flux relative-flux contours for the noise added deconvolved $20.67 < J < 21.2$ contaminant model shown in Fig. 6 (red), compared to the real (noisy) data distribution (black). Errors have been added as explained in Appendix B. The labelled quantities are relative fluxes (i.e. fluxes in different bands divided by the *J*-band flux). It is apparent that, after re-adding the errors to the deconvolved quantities, we obtain a distribution that is consistent with the $20.67 < J < 21.2$ real data.

5.1 Completeness and efficiency computation

To select high-z QSO candidates for spectroscopic follow-up confirmation, we defined a probability threshold ($P_{\rm th}$) that effects a balance between contamination and completeness: this threshold should be small enough to avoid missing many high-z QSOs, so that the sample completeness is high, and it should be large enough to keep the number of contaminants low to increase the efficiency of the selection method. From a practical perspective, the completeness can be seen as a proxy for the expected fraction of recovered high-z QSOs as a function of the probability threshold, in a certain sky area, while the efficiency is a proxy for the expected spectroscopic confirmation efficiency of the candidates at the telescope.

The completeness (C) is defined as:

$$C \equiv \frac{N_{\rm Q}(P \geq P_{\rm th})}{N_{\rm Qtot}}, \tag{10}$$

where $N_{\rm Q}(P \geq P_{\rm th})$ is the number of high-z QSOs per square degree with a probability $P \geq P_{\rm th}$, and $N_{\rm Qtot}$ is the total number of QSOs per square degree down to $J \leq 21.5$, while the efficiency (E) is defined as:

$$E \equiv \frac{N_{\rm Q}(P \geq P_{\rm th})}{N_{\rm Q}(P \geq P_{\rm th}) + N_{\rm C}(P \geq P_{\rm th})}, \tag{11}$$

where $N_{\rm C}(P \geq P_{\rm th})$ is the number of contaminants with a probability $P \geq P_{\rm th}$ per square degree. In the limit where the classification of all

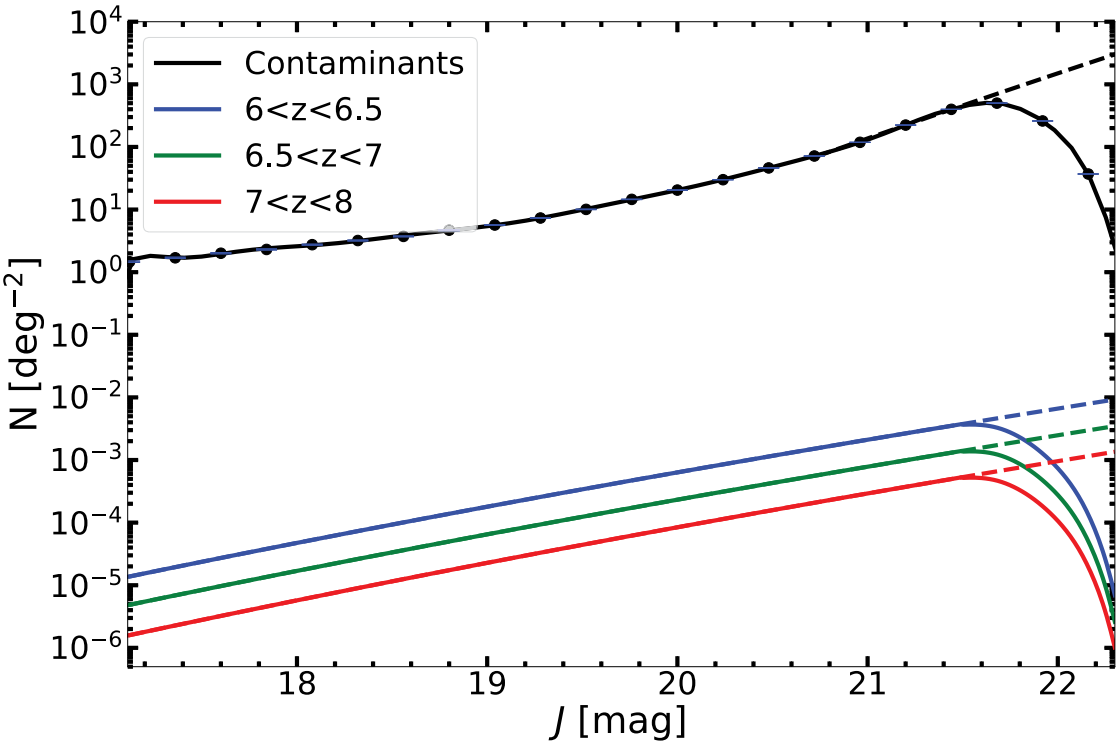

Figure 8. Number counts $p(\hat{F}_J|O \in \text{“}class''\text{)}$ priors for the contaminant (black line and points), and the $6 \leq z \leq 6.5$ (blue line), $6.5 \leq z \leq 7$ (green line), $7 \leq z \leq 8$ (red line) QSO classes as a function of the J-band magnitude. The black points are the real contaminant data from the VIKING survey, while we used a 40-order polynomial to perform a robust fit to the range $J \leq 21.4$, and at $J > 21.4$ we used a cubic spline to interpolate the histogram, namely to capture the drop-off due to catalogue incompleteness (black line). To model the effect of the incompleteness on the real data distribution, we fit a power law to the range $20.7 \leq J \leq 21.4$, and extrapolated it to $J > 21.4$ (black dashed line). The ratio between the value given by the power law and the cubic spline interpolated number counts gives us the incompleteness correction term to apply to our QSO number counts at $J > 21.4$. The 1σ Poissonian errors are shown as short blue lines. The other three QSO colored lines show the $z \sim 6.7$ quasar LF from Wang et al. (2019), after the inclusion of the incompleteness. The corresponding extrapolation of the LF at $J > 21.4$ without the incompleteness correction is shown as a dashed line.

the sources in our survey is known, we could compute both C and E directly from the VIKING survey area. However, as we do not know the real classification of most of the sources in our sample, we used simulations to compute the completeness and the efficiency of our selection method, as we now describe.

In order to reduce the statistical fluctuations we simulated a large number of both high-z QSOs and contaminants. High-z QSOs were simulated by sampling the $z \geq 6$ LF from equation (7) (Wang et al. 2019), using a Monte Carlo Simulation (MCS) approach. Namely, this equation can be interpreted as the 2D probability distribution of the quasars as a function of redshift and magnitude, and MCS is a convenient method to generate samples. Again, we expressed the LF as a function of redshift and apparent J-band magnitude, by converting the M_{1450} to J-band magnitude using the k-correction from equation (9), and multiplied it by the incompleteness found in Section 4.2 for the VIKING J-band magnitude distribution. We then used the MCRS method to sample the redshift and J-band magnitude distributions of 400 000 QSOs with $6 \leq z \leq 8$, and $17 \leq m_J \leq 22$. Given the redshift and J-band magnitude of each source, we used our deconvolved quasar models to sample the noiseless fluxes for the 300 000 simulated QSOs, and added representative photometric errors according to our noise model in Appendix B. Then, the simulated QSOs were divided into the three redshift bins adopted previously, and we computed their probability of being quasars using equation (1), to derive the $N_{\rm Q}(P \geq P_{\rm th})$ needed for equations (10) and (11).

To simulate the contaminants, we drew 100 million $17 \leq m_J \leq 22$ sources from the J-band magnitude distribution of the contaminant training catalog (upper-left panel Fig. 8). We again sampled the deconvolved contaminant models to generate the noiseless fluxes for our simulated sources, and added the errors as explained in Appendix B. Then, we evaluated the probability that these synthetic sampled ‘sky’ objects are quasars using equation (1), which is needed to determine the $N_{\rm C}(P \geq P_{\rm th})$ term from equation (11). Finally, we rescaled the numbers of simulated contaminants and high-z QSOs to reflect the prior number count distributions shown in Fig. 8, and we used equations (10) and (11) to compute the completeness and efficiency, down to a J-band magnitude of 21.5. This magnitude limit was introduced since it is representative of what can be realistically confirmed with a near-IR instrument on an 8-m class telescope in a reasonable exposure time, and is also close to the 5σ limit of the VIKING data we use. Fainter objects would require longer exposure times and excellent observing conditions making them much more challenging to spectroscopically confirm.

In Fig. 10, we display the number count distribution of quasar probabilities, $\mathrm{d}N/(\mathrm{d}\Omega/\mathrm{d}P)$, for simulated QSOs and contaminants. This quantity is defined such that the integral over probability P yields the number of objects per square degree. Fig. 11 shows the efficiency (black) and the completeness (red) of our selection method as a function of the probability threshold ($P_{\rm th}$), in the three redshift bins: $6 \leq z \leq 6.5$ (top), $6.5 \leq z \leq 7$ (central), and $7 \leq z \leq 8$ (bottom). It is apparent that lowering the threshold will

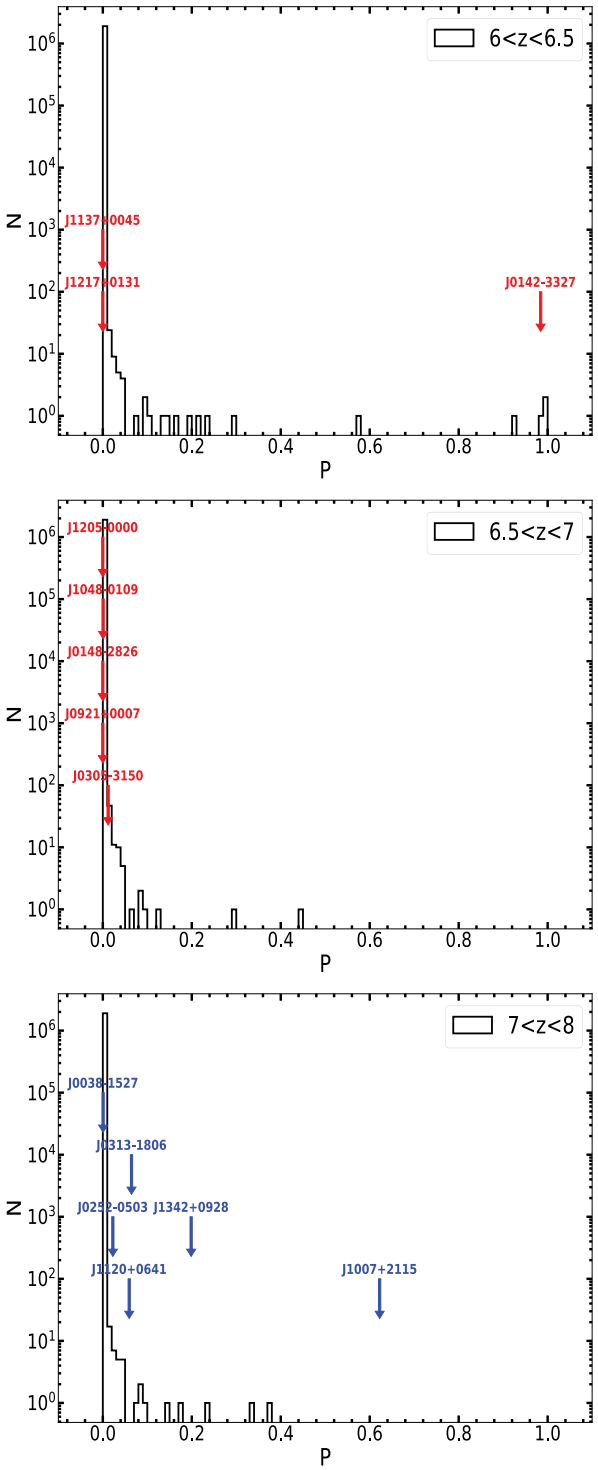

Figure 9. Probability distributions of sources from our VIKING candidate catalogue in three different redshift ranges: $6 \leq z \leq 6.5$ (top), $6.5 \leq z \leq 7$ (central), and $7 \leq z \leq 8$ (bottom). This catalogue has also been used to train the contaminant models, as most of these sources are expected to be contaminants. The three distributions are obtained by doing model comparison between the contaminant model and, separately, each of the three high-redshift quasar models, as explained in Section 4.1. The downward red arrows highlight the probability of known high-z QSOs in the VIKING survey area. Candidates with $P \sim 0$ are pinpointed with arrows plotted on top of each other. In the bottom panel, downward blue arrows highlight the probability of known $z > 7$ QSOs in the entire sky.

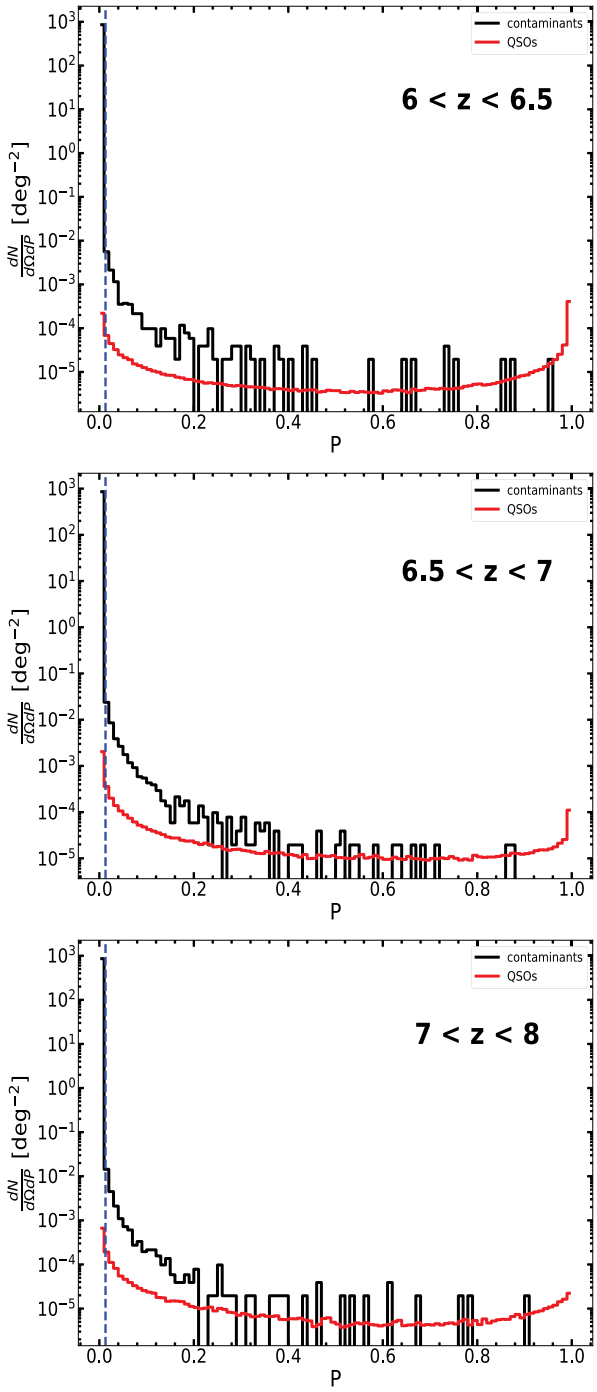

Figure 10. Probability distributions of simulated contaminants (black) and high-z QSOs (red) per square degrees, in three different redshift ranges: $6 \leq z \leq 6.5$ (top), $6.5 \leq z \leq 7$ (central), and $7 \leq z \leq 8$ (bottom). The blue dashed vertical line marks our adopted probability threshold.

always increase the completeness, but this comes at the cost of a lower efficiency, thus increasing the number of contaminants that are spectroscopically followed up. It is also evident that the completeness and efficiency are generally higher in the $6 \leq z \leq 6.5$ range, where the overlap between the QSO and contaminant relative-flux distributions is smaller compared to the $6.5 \leq z \leq 7$, and $7 \leq z \leq 8$ cases (i.e. the red and green contours overlap the black contours in Fig. 5 more than the blue contours).

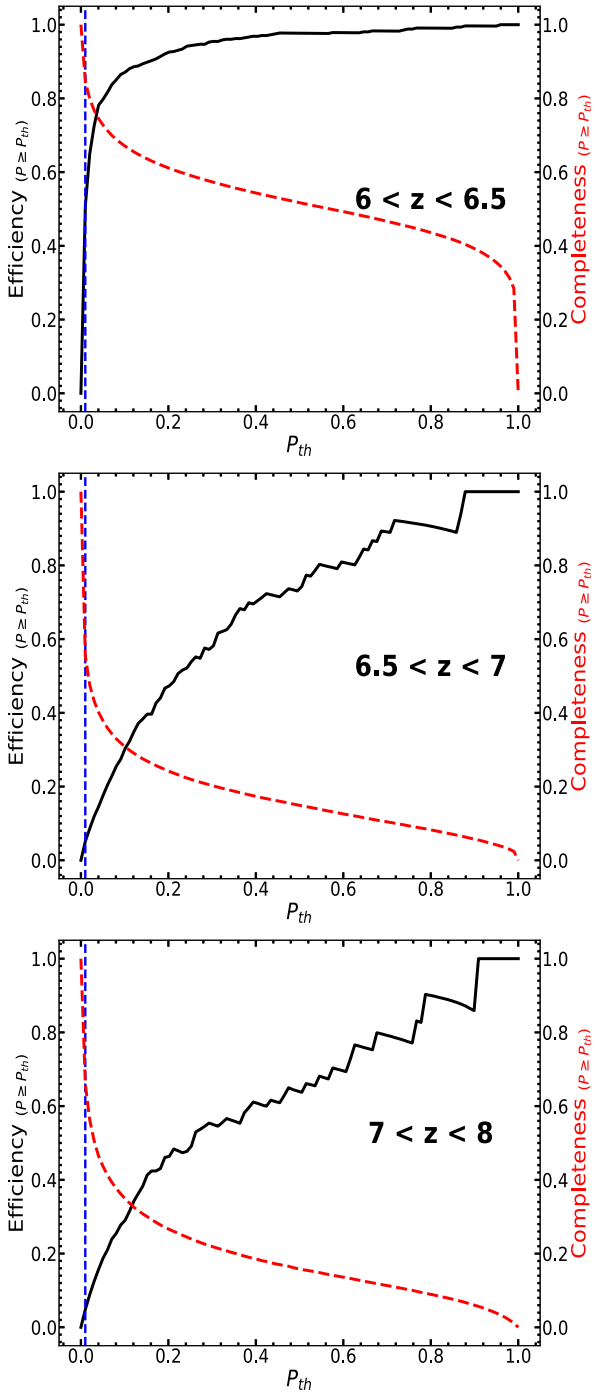

Figure 11. Efficiency (black solid line) and completeness (red dashed line) of the XDHZQSO selection method as a function of probability threshold ($P_{\rm th} = 0.01$). The three panels show the results for the three redshift bins: $6 \leq z \leq 6.5$ (top), $6.5 \leq z \leq 7$ (central), and $7 \leq z \leq 8$ (bottom). The blue dashed vertical line marks our adopted probability threshold. It is apparent that lowering the threshold will always increase the incompleteness but this comes at the cost of lower efficiency, thus increasing the number of contaminants selected for spectroscopic follow-up. It is also evident that both the efficiency and completeness are lower at $6.5 \leq z \leq 7$, and $7 \leq z \leq 8$, where the QSO properties largely overlap with the contaminant distribution (see the overlap between the red and green over the black contours in Fig. 5).

Since the expected number density of high-z QSOs is very low, the choice of the $P_{\rm th}$ is mostly determined by the need to have a high completeness to avoid missing the coveted highest redshift sources. In fact, by integrating the LF in equation (7) down to $J = 21.5$, we expect to find ≈15, ≈5, and ≈2 QSOs in the ranges $6 \leq z \leq 6.5$, $6.5 \leq z \leq 7$, and $7 \leq z \leq 8$, respectively, in the 1076 deg^2 VIKING survey area. Recovering this small number of expected sources would require a relatively high completeness (possibly $C \approx 85$ per cent). As such, we chose to use the completeness as the main criterion for setting the probability threshold $P_{\rm th}$, whereas the efficiency plays a pivotal role in setting $P_{\rm th}$ when a high completeness corresponds to $E < 5$ per cent. To visualize the trade-off between completeness and efficiency (both of which are parametrized by $P_{\rm th}$), we plot in Fig. 12 the efficiency as a function of the completeness for the three redshift bins.

In the $6 \leq z \leq 6.5$ range, the 85 per cent completeness requirement corresponds to $P_{\rm th} = 0.01$ and $E = 50$ per cent (Fig. 12, top panel). In the $6.5 \leq z \leq 7$ range the high completeness requirement ($C = 85$ per cent) cannot be achieved without lowering the efficiency to an unacceptable value ($E \approx 10^{-3}$ per cent; see Fig. 12, central panel), while a 56 per cent completeness (achievable with $P_{\rm th} = 0.01$) corresponds to $E \approx 5$ per cent, which is a more reasonable efficiency value to work with. For the $7 \leq z \leq 8$ range, the same probability threshold adopted for the lower redshift bins ($P_{\rm th} = 0.01$) provides a completeness of 66 per cent and an efficiency of 5 per cent. The very low value of efficiency in the two highest redshift ranges is caused by the large overlap between the $6.5 \leq z \leq 8$ QSOs and the contaminant models, as is apparent in Fig. 5 (see the larger overlap of the green and red with the black contours). Consequently, also the number of QSO candidates with probability above the threshold in these redshift ranges is lower compared to the $6 \leq z \leq 6.5$ range.

To summarize, we report in Table 3 the three probability thresholds derived from our completeness and efficiency analysis, and the corresponding completeness, efficiency, and number of candidates $N(P_{\rm QSO} \geq P_{\rm th})$ with $P_{\rm QSO} \geq P_{\rm th}$ that are selected for future spectroscopic follow-up. For the $7 \leq z \leq 8$ range, we obtain an efficiency that is 5 per cent, whereas quasar selections based on colour-cuts work at a lower level efficiency in this redshift range (Bañados et al. 2018; Wang et al. 2021). The higher efficiency that we derive results from the combination of two primary factors: (1) our probabilistic density estimation takes advantage of the full feature space (all flux ratios) at once without strict boundaries, making it more effective and inclusive than simple colour-cuts, and (2) our effort to compile as much panchromatic photometry as possible improves the efficiency, relative to previous efforts (Mortlock et al. 2011; Bañados et al. 2018; Yang et al. 2020a; Wang et al. 2021), to select $z > 7$ quasars using colour cuts. On the other hand, an efficiency of 5 per cent for the $6.5 \leq z \leq 7$ range is lower compared to some colour-cut selections performed in the past (i.e. Bañados et al. 2016). This likely results from the fact that our study does not include the PS1-zy filters, which greatly improves the selection of $6.5 \leq z \leq 7$ QSOs, since in this redshift range the Ly α line enters the PS1-y filter and drops out of the PS1-z filter, while the broader DECaLS-z filter covers both the aforementioned PS1 filters. In a future study, we plan to include the PS1-zy filters to improve our selection efficiency for this particular redshift range.

6 CLASSIFICATION OF KNOWN HIGH-z QUASARS

By integrating the $z = 6.7$ LF from Wang et al. (2017) in the $17 \leq J \leq 21.5$ range, we expect to find ≈21 (≈28) QSOs at $6 \leq z \leq$

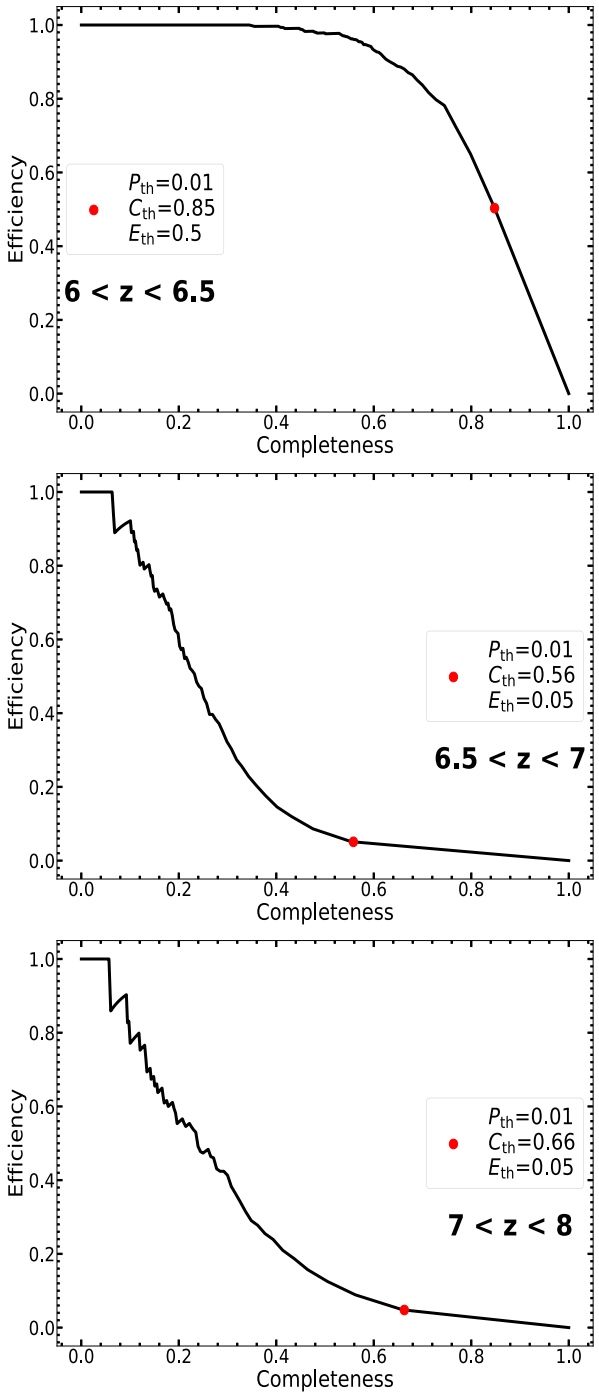

Figure 12. Efficiency versus completeness in the three redshift bins: $6 \leq z \leq 6.5$ (top), $6.5 \leq z \leq 7$ (central), and $7 \leq z \leq 8$ (bottom). The red point marks the efficiency and completeness at the value of the chosen probability threshold (P_{th}). The low overlap between the $6 \leq z \leq 6.5$ QSO and contaminant contours allows us to work with high values of efficiency (68 per cent) and completeness (90 per cent). However, at $6.5 \leq z \leq 7$ and $7 \leq z \leq 8$ the overlap with the contaminant properties is so large that we are forced to work at a lower efficiency (10 per cent) to have a high completeness ($\geq$ 75 per cent).

6.5, ≈7 (≈9) QSOs at $6.5 \leq z \leq 7$, and ≈3 (≈4) QSOs at $7 \leq z \leq 8$, depending on whether (or not) we consider the effect of the *J*-band photometric incompleteness in the VIKING survey. Thus, after performing the spectroscopic follow-up of the targets with $P \geq P_{th}$, we expect to discover high-z QSOs among our candidates with numbers consistent with these estimates.

Past works already studied the VIKING area and searched for $z \geq 6$ QSOs (e.g. Venemans et al. 2013, 2015; Barnett et al. 2021). For example, both Venemans et al. (2013) and Barnett et al. (2021) used the $ZYJHK_s$ filters from the VIKING survey to find $z > 6.5$ QSOs: Venemans et al. (2013) applied colour-cuts and found three new QSOs, while Barnett et al. (2021) selected four known QSOs and 17 QSO candidates using the BMC method, but no new QSOs were found. Other QSOs were found in the VIKING footprint from past works, as they searched for high-z QSOs in other surveys that partially overlap with the VIKING area: i.e. the CFHTLS (Willott et al. 2009), the Pan-STARRS1 (Bañados et al. 2016), the VST-ATLAS (Carnall et al. 2015), the DELS (Wang et al. 2017), and the HSC (Matsuoka et al. 2016, 2018a, b, c, 2019a, b). So, we expect to have some known high-z QSOs in our VIKING data set, and to recover them among our candidates. In Section 6.1, we provide a summary of the known QSOs that are covered within our search area but that are not in our VIKING data set due to our selection criteria. Then, we describe the performance of XDHZQSO in recovering and classifying both the known high-z QSOs in the VIKING survey area (Section 6.2), as well as the known $z > 7$ QSOs (Section 6.3) over the entire sky.

6.1 Missed high-z QSOs

From past works (Willott et al. 2009; Venemans et al. 2013; Bañados et al. 2016; Matsuoka et al. 2016, 2018a, b, c, 2019a, b), we identified 32 known $z \geq 6$ QSOs in the DECaLS + VIKING area. However, the imposition of our selection criteria reduced this number in our final VIKING area data set, as 20 QSOs are lost because they do not satisfy SNR(J) > 5, and another four QSOs are not selected as they do not have data in all the bands considered in our study. That leaves eight known $z > 6$ quasars in the VIKING area data set whose probabilistic classification is described in the following section.

6.2 Classification of known high-z QSOs in the VIKING Survey Area

Among the classified sources there are eight known high-z QSOs that were found in the VIKING survey area from past works: DELS J1217 + 0131 ($z = 6.17$; Bañados et al. 2016; Wang et al. 2017), ATLAS J025.6821-33.4627 ($z = 6.31$, hereafter J0142−3327; Carnall et al. 2015), HSC J1137+0045 ($z = 6.4$; Matsuoka et al. 2019b), J0148−2826 ($z = 6.54$; Yang et al. 2020b), HSC J0921+0007 ($z = 6.56$; Matsuoka et al. 2018b), VIK J0305−3400 ($z = 6.604$; Venemans et al. 2013), DELS J1048−0109 ($z = 6.63$; Wang et al. 2017), HSC J1205−0000 ($z = 6.74$; Matsuoka et al. 2016). These sources and their main properties are listed in Table 4, while their probabilities of being high-z QSOs are pinpointed with red arrows in Fig. 9.

In the range $6 \leq z \leq 6.5$, our models are able to correctly classify one known QSO out of three, J0142−3327 ($P_{QSO} \approx 98.4$ per cent), but our selection threshold in this redshift range ($P_{th} = 1$ per cent) does not allow us to recover DELS J1217+0131 ($P_{QSO} \approx 3 \times 10^{-6}$ per cent), and HSC J1137+0045 ($P_{QSO} \approx 4 \times 10^{-9}$ per cent). The probability of these three quasars are also reported in Table 4 and shown in Fig. 9 (upper panel). The low probability of the

Table 3. Number of selected candidates in the three redshift bins.

z range	$P_{\rm th}$ (per cent)	$C_{\rm th}$ (per cent)	$E_{\rm th}$ (per cent)	$N(P_{\rm QSO} \geq P_{\rm th})$	$N_{\rm exp}$	$N_{\rm rec}$
6.0–6.5	1	85	50	58	15	10
6.5–7.0	1	56	5	80	5	1
7.0–8.0	1	66	5	43	2	2

Note. Summary of the probability threshold ($P_{\rm th}$) adopted in each redshift bin to select high-z QSO candidates for spectroscopic follow-up, and the corresponding completeness ($C_{\rm th}$), efficiency ($E_{\rm th}$), and number of candidates selected ($N(P_{\rm QSO} \geq P_{\rm th})$). The last two columns represent the number of QSOs expected ($N_{\rm exp}$) according to our adopted LF (equation (7)) down to $J = 21.5$, and how many of them we expect to recover among our candidates ($N_{\rm rec}$).

Table 4. Known QSOs in the VIKING survey area.

Name	z	J	$P_{\rm QSO}$	Ref.
DELS J1217+0131	6.17	21.28 ± 0.14	3×10^{-6} per cent	Bañados et al. (2016); Wang et al. (2017)
ATLAS J025.6821−33.4627	6.31	19.02 ± 0.02	98.4 per cent	Carnall et al. (2015)
HSC J1137+0045	6.4	21.51 ± 0.20	4×10^{-9} per cent	Matsuoka et al. (2019b)
J0148−2826	6.54	21.09 ± 0.13	4×10^{-3} per cent	Yang et al. (2020b)
HSC J0921+0007	6.56	20.9 ± 0.26	2×10^{-7} per cent	Matsuoka et al. (2018b)
VIK J0305−3400	6.61	20.07 ± 0.09	1.2 per cent	Venemans et al. (2013)
DELS J1048−0109	6.63	20.99 ± 0.12	0.07 per cent	Wang et al. (2017)
HSC J1205−0000	6.75	21.95 ± 0.21	3×10^{-12} per cent	Matsuoka et al. (2016)

latter one is not surprising, considering that HSC J1137+0045 is a very faint QSO ($J = 21.51$ and SNR(J) = 5.4), selected from the Hyper Suprime-Cam (HSC) Subaru Strategic Program (SSP) survey (Aihara et al. 2018), and that apparently lacks strong Ly α in emission (Matsuoka et al. 2019b). However, to better understand the low probability values obtained for these two quasars, we compared their photometric properties with those sampled from our XD deconvolved models. For each of the three known $6 \leq z \leq 6.5$ QSOs, we simulated 10 000 contaminants and 10 000 $6 \leq z \leq 6.5$ QSOs, using the XDHZQSO models in the magnitude bins that include the J-band magnitudes of the three QSOs. To visualize the probability of selecting a known quasar, we draw samples from the 'deconvolved' (i.e. noise free) XDHZQSO contaminant and quasar models, and overplot the relative flux measurements of the real quasars, with ellipses indicating their (covariant) 1σ errors. This is shown in Fig. 13, where we plot the deconvolved relative-flux relative-flux contours for the simulated contaminants (black) and $6 \leq z \leq 6.5$ QSOs (blue), compared to the properties of the known $6 \leq z \leq 6.5$ QSOs from the VIKING survey area. The reason we are creating 10 000 copies of contaminant and 10 000 of QSOs for each known high-z QSO is that the contaminant and quasar models are magnitude dependent. Thus formally, we would need to show a plot for each object, where we compare its properties with those from the sampled contaminants and QSOs. However, given that these magnitude dependencies are subtle, we chose to simply simulate 10 000 copies of sources at each magnitude and aggregate them on to a single plot. It is apparent that in some sub-plots of Fig. 13 (especially those with f_z/f_J and f_Y/f_J), the relative fluxes of both HSC J1137+0045 and DELS J1217+0131 are not consistent with the simulated $6 \leq z \leq 6.5$ QSOs relative flux distributions (blue contours), consequently lowering the classification probability of these two objects. Considering that HSC J1137+0045 is a QSO that apparently lacks strong Ly α in emission (Matsuoka et al. 2019a), while DELS J1217+0131 exhibits a strong Ly α emission line (Wang et al. 2017), we conclude that the properties of the 'simqso' simulated high-z QSOs, that have been used for the training of our XDHZQSO QSO models, are too rigid to include these two sources.

In the range $6.5 \leq z \leq 7$, as reported in Table 4 and displayed in Fig. 9 (middle panel), our method is able to recover one QSO (based on our $P_{\rm th} = 1$ per cent), VIK J0305−3400 ($P_{\rm QSO} \approx 1.2$ per cent), while the other four are consistent with being contaminants ($P_{\rm QSO} \leq 10^{-1}$ per cent). Among them, J0921+0007 ($P_{\rm QSO} \approx 2 \times 10^{-7}$ per cent) is also an HSC selected QSO ($J = 20.9$) that has similar optical colors to Galactic brown dwarfs (Matsuoka et al. 2018b). Adopting the same procedure as described above to generate 10 000 contaminants and $6.5 \leq z \leq 7$ QSOs for each known QSO, we show in Fig. 14 the deconvolved relative-flux relative-flux contours for the simulated contaminants (black) and high-z QSOs (blue), compared to the properties of the known $6.5 \leq z \leq 7$ QSOs from the VIKING survey area. Also in this case, it is apparent that the relative fluxes of the four QSOs with $P_{\rm QSO} \leq 10^{-1}$ per cent are inconsistent with the deconvolved QSO model properties (blue contours in Fig. 14) in some sub-plots: (1) J0148−2826 is inconsistent with panels showing f_H, f_{W1}, and f_{W2}, (2) HSC J0921+0007 is inconsistent with panels showing f_{W1}, and f_{W2}, (3) DELS J1048−0109 is not consistent with panels showing f_H, and f_{W2}, and (4) HSC J1205−0000 is not consistent with the QSO distribution in any panel. We provide a more detailed discussion of these discrepancies between real and simulated QSO properties in Section 7.1.

6.3 Classification of the z ≥ 7 QSOs

While we tested in Section 6.2, the ability of our models to recover the known $6 \leq z \leq 7$ QSOs in the VIKING survey area, testing our classification models for the highest redshift range was not possible as there are no known $z > 7$ QSOs in the VIKING footprint. Therefore, we applied our method to the $z > 7$ QSOs that have been discovered so far over the entire sky, using published photometric measurements.

There are, at the time of writing, a total of eight known $z > 7$ QSOs: J2356+0017 ($z = 7.01$; Matsuoka et al. 2019b), J0252−0503 ($z = 7.02$; Yang et al. 2019), J0038−1527 ($z = 7.021$; Wang et al. 2018), J1243+0100 ($z = 7.07$; Matsuoka et al. 2019a), J1120+0641 ($z = 7.085$; Mortlock et al. 2011), J1007+2115 ($z = 7.515$; Yang et al. 2020a), J1342+0928 ($z = 7.541$; Bañados et al. 2018), and J0313−1806 ($z = 7.642$; Wang et al. 2021). To classify them, we

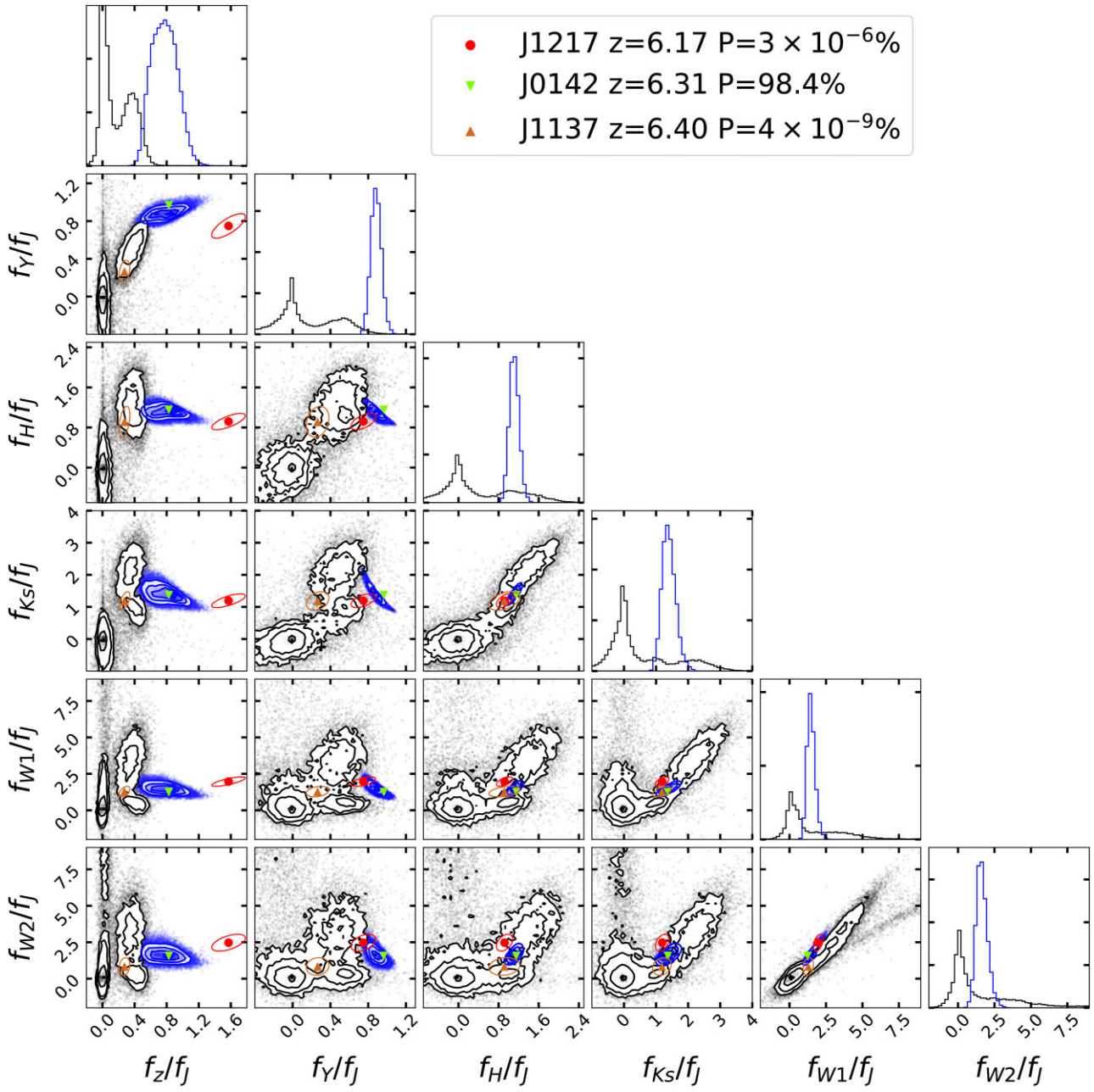

Figure 13. Deconvolved relative-flux relative-flux contours for the simulated contaminants (black) and $6 \leq z \leq 6.5$ QSOs (blue), compared to the real (noisy) properties of the known $6 \leq z \leq 6.5$ QSOs from the VIKING survey area. The probability threshold to select these sources with our method is $P_{\rm th} = 0.01$. It is apparent that both J1217 and J1137 are 'off' from the QSO contours in the f_z/f_J sub-plots, while J1137 is also 'off' in the f_Y/f_J sub-plots, thus lowering their probabilities of being classified as high-z QSOs.

first collected the photometric data in the seven bands of interest (DECaLS-z, VIKING-$YJHK_S$, and *WISE*-$W1W2$) from the literature, when available. Since some of these sources have public NIR data coming from the Wide Field Infrared Camera (WFCAM) for the UK Infrared Telescope (UKIRT), we used the transformation equations between VISTA and WFCAM derived by González-Fernández et al. (2018), to convert the UKIRT magnitudes into the VIKING ones. For the missing flux measurements, we performed forced photometry. Since J0313−1806 has no photometric measurements in the Y and H bands, we used synthetic photometry computed by integrating the observed spectrum of this source from Wang et al. (2021) against the respective filter curves. However, we excluded from our classification list both J2356+0017 and J1243+0100, as they are too faint (SNR(J) < 5) to make it into our catalog. Finally, we used our XDHZQSO models to classify the remaining six sources following the same procedure described in Section 5. In Table 5, we summarize the properties and results from our classification of these six $z \geq 7$ QSOs.

Based on our defined probability threshold for the $z \geq 7$ range ($P_{\rm th} = 1$ per cent), we are able to recover five QSOs: J0252−0503 ($P_{\rm QSO} = 2.3$ per cent), J1120+0641 ($P_{\rm QSO} = 5.9$ per cent), J1007+2115 ($P_{\rm QSO} = 62.2$ per cent), J1342+0928 ($P_{\rm QSO} = 19.9$ per cent), and J0103−1806 ($P_{\rm QSO} = 6.5$ per cent). However, we fail to select J0038−1527 ($P_{\rm QSO} = 0.07$ per cent).

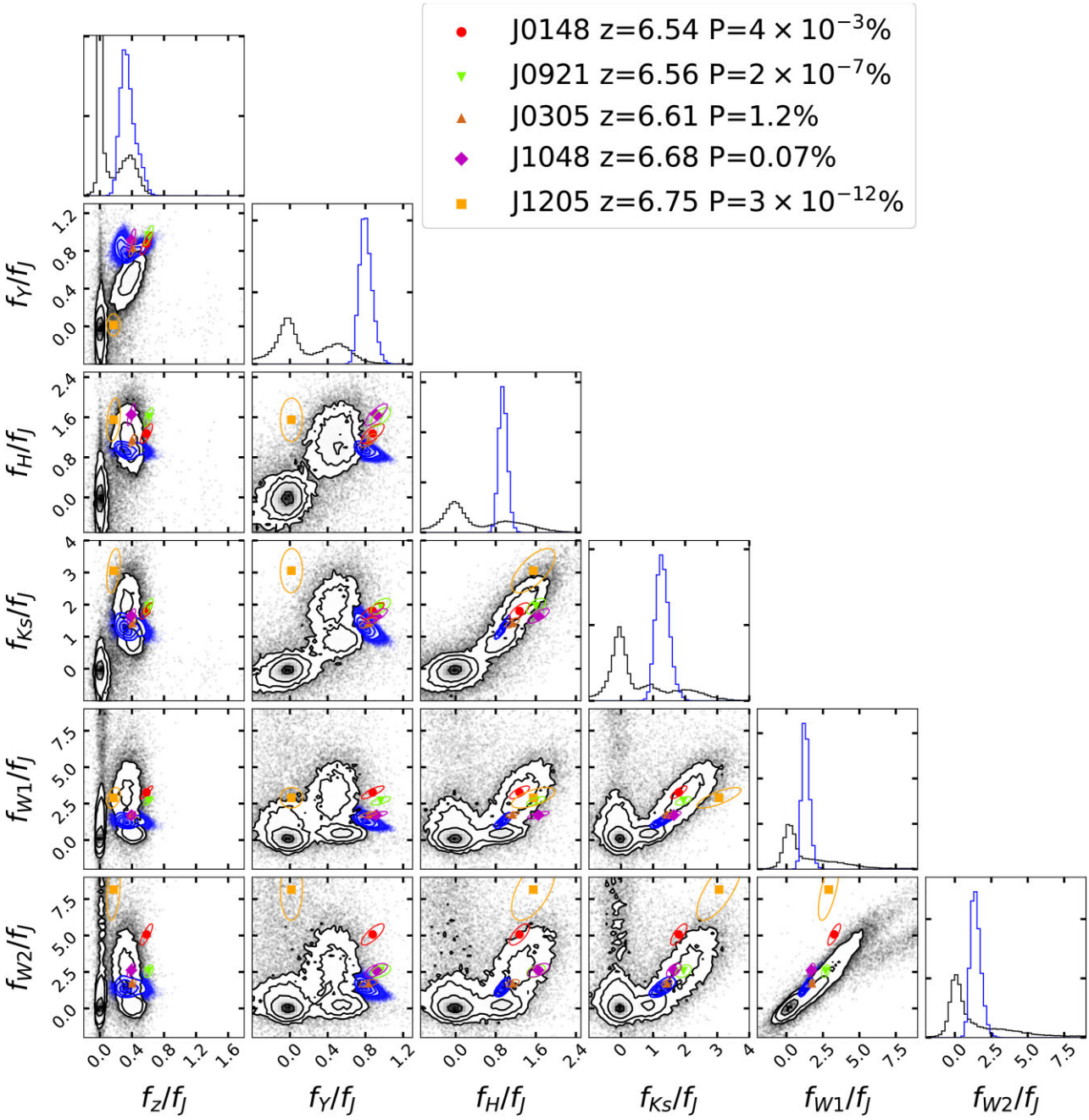

Figure 14. Same as Fig. 13 but in the $6.5 \leq z \leq 7$ bin. The probability threshold to select these sources with our method is $P_{\rm th} = 0.01$. The four QSOs with $P_{\rm QSO} \leq 10^{-2}$ per cent are inconsistent with the deconvolved QSO model properties (blue contours) in the following sub-plots: (1) J0148 is inconsistent with panels showing f_H, f_{W1}, and f_{W2}, (2) J0921 is inconsistent with panels showing f_{W1}, and f_{W2}, (3) J1048 is not consistent with panels showing f_H, and f_{W2}, and (4) J1205 is not consistent with the QSO distributions in any panel.

Table 5. Known $z \geq 7$ QSOs classified by our XDHZQSO method.

Name	z	J	$P_{\rm QSO}$	Ref.
J0252−0503	7.02	21.13 ± 0.07	2.3 per cent	Yang et al. (2019)
J0038−1527	7.021	20.63 ± 0.08	0.07 per cent	Wang et al. (2018)
J1120+0641	7.085	21.22 ± 0.17	5.9 per cent	Mortlock et al. (2011)
J1007+2115	7.515	21.14 ± 0.18	62.2 per cent	Yang et al. (2020b)
J1342+0928	7.541	21.24 ± 0.02	19.9 per cent	Bañados et al. (2018)
J0313−1806	7.642	20.92 ± 0.13	6.5 per cent	Wang et al. (2021)

J0038−1527 exhibits strong broad absorption line (BAL) features (Wang et al. 2018), that can alter its colors, making it different compared to our $7 \leq z \leq 8$ QSO models, which do not attempt to model BAL absorption. As in Section 6.2, we simulated a large number of contaminants and $7 \leq z \leq 8$ QSOs, and compare their relative fluxes with those from the real $z > 7$ QSOs in Fig. 15. It is evident that J0038−1527 deviates from the blue contours (deconvolved $7 \leq z \leq 8$ QSO models) in the sub-plot displaying f_z/f_J versus f_Y/f_J, as the absorption from the BALs impacts the Y-band

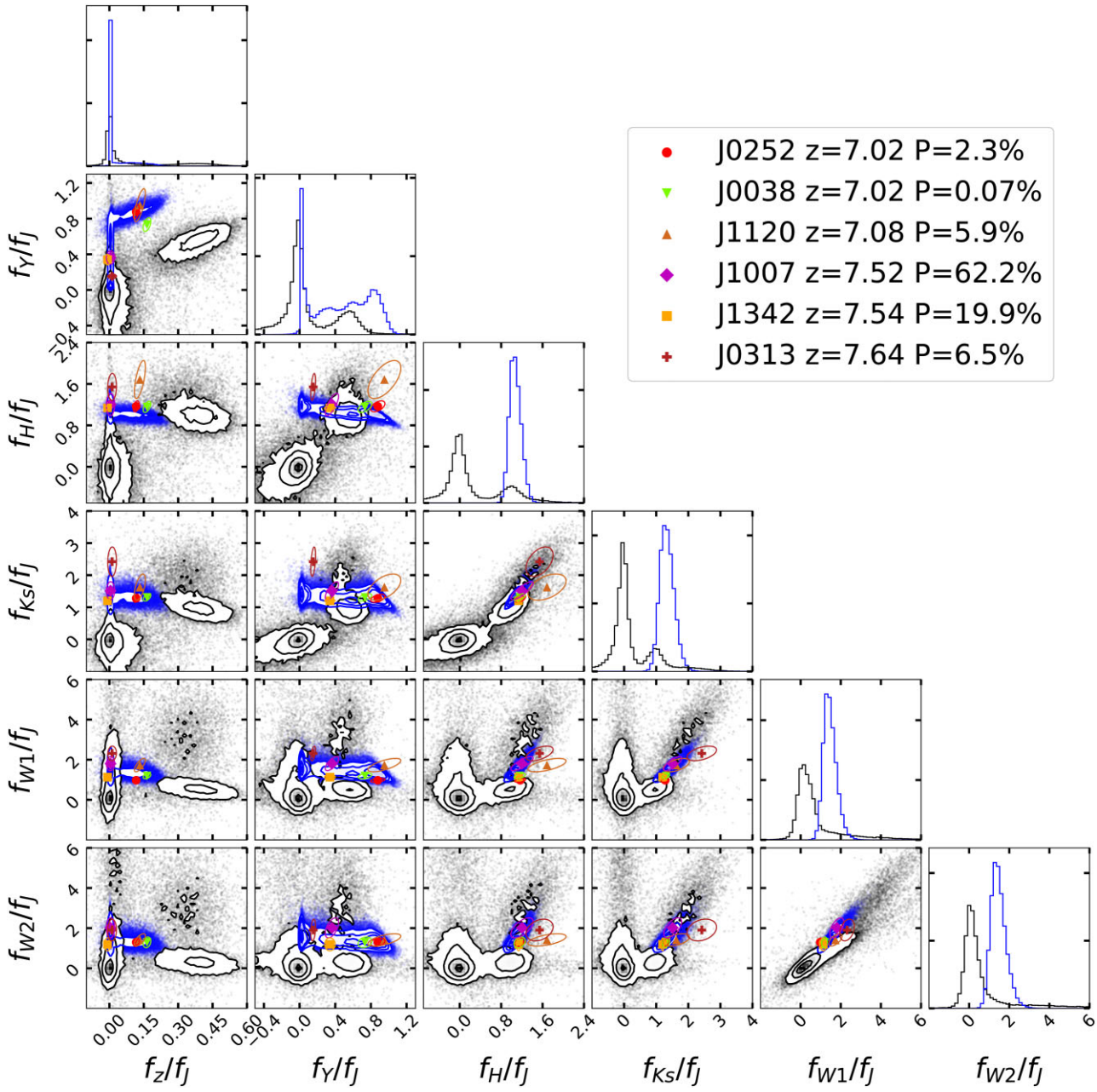

Figure 15. Deconvolved relative-flux relative-flux contours for the simulated contaminants (black) and $7 \leq z \leq 8$ QSOs (blue), compared to the real (noisy) properties of the known $z \geq 7$ QSOs to date. The probability threshold to select these sources with our method is $P_{\rm th} = 0.01$. It is apparent that J0038−1527 deviates from the quasar locus indicated by the blue contours in the sub-plot displaying f_z/f_J versus f_Y/f_J, with the effect of lowering its QSO classification probability.

flux. Again, we discuss the deviations of the real QSO properties from the expected simulated ones in Section 7.1.

7 DISCUSSION

In Section 6, we showed that XDHZQSO is able to recover two $6 \leq z \leq 7$ QSOs out of eight that passed our selection criteria (i.e. SNR$(J) \geq 5$, SNR$(g, r) < 3$, and no missing data) and made into our catalogue: we select one QSO at $6 \leq z \leq 6.5$ (ATLAS J025.6821−33.462 at $z = 6.31$), and one QSO at $6.5 \leq z \leq 7$ (VIK 0305−3400 at $z = 6.61$). The application of XDHZQSO on the $7 \leq z \leq 8$ QSOs found in the entire sky, that meet our selection criteria, allows us to recover five out of six $7 \leq z \leq 8$ QSOs (J0252−0503 at $z = 7.02$, 1120+0641 at $z = 7.085$, J1007+2115 at $z = 7.515$, J1342+0928 at $z = 7.541$, and J0313−1806 at $z = 7.642$). In Section 7.1, we discuss the limitations of our selection technique that could explain our failure to select of some of the known high-z QSOs in the VIKING area, while in Section 7.2 we provide a comparison between our code and other probabilistic classification methods.

7.1 Limitations of the XDHZQSO selection method

In Sections 6.2 and 6.3, we showed that our method is only able to recover some of the known high-z QSOs. In fact, there are several

reasons that can lead to the failure to select a source, and all of them involve the source properties and corresponding errors being more consistent with the XDHZQSO contaminant models rather than the high-z QSO ones. Here we discuss the possible causes that lead to the non-selection of some of the known $z > 6$ sources:

(i) *Noisy data.* In the case of a source with large photometric errors, our method naturally degrades its probability of belonging to high-z QSOs class if the data uncertainties imply that the object overlaps with the contaminant class. On the other hand, this limitation is not afflicting other selection methods. In fact, a colour-selection technique that does not use photometric errors could select a noisy object, whereas XD would spread that probability out, meaning it might be more likely to be classified as a contaminant if, given the errors, it significantly overlaps the contaminant locus. However, we stress that taking errors into account is a feature not a flaw of our method (i.e. not taking into account errors will generally result in an overall lower efficiency then taking them into account).

(ii) *Photometric variability.* Since the surveys considered in this work were performed at different epochs, intrinsic variability of sources could also play a role in lowering the computed probabilities (see Ross & Cross 2020 for a study of the variability of $5 < z < 7$ quasars). However, since the variability of these objects is supposed to be small (at most 10 per cent given low-z structure functions; e.g. Vanden Berk et al. 2004; Kelly, Bechtold & Siemiginowska 2009; Schmidt et al. 2010), we argue that this is probably not the main issue we are facing.

(iii) *Inaccurate models.* Since our method is a classification technique, its validity strongly depends on the correct modelling of the considered classes. If the XDHZQSO models are not a good representation of the underlying deconvolved flux distributions of one or more classes, then the computed probabilities are not reliable. Although, that seems not the case for our contaminant class, as the models are trained with the real data coming from our survey, it can be an issue for our high-z QSO classes. In fact, our quasar models are trained on synthetic photometry determined from simulated QSO spectra whose properties are consistent with the mean spectrum of low-z luminous QSOs (McGreer et al. 2013). However, these simulated quasar spectra could not well represent the intrinsic relative flux scatter of all the luminous QSOs, or the properties of peculiar sources such as Broad Absorption Line QSOs (BALQSOs). For example, J0038−1527 is a BALQSO (Wang et al. 2018), and its Y-band relative flux is lower than expected compared to objects with similar redshift and luminosity (see Fig. 15). Furthermore, in the sub-panels showing H, K, $W1$, and $W2$ bands in Figs 13, 14, 15 it is apparent that our XDHZQSO QSO models are too rigid, as the simulated QSO deconvolved density distributions (blue contours) appear too little scatter as compared to the real QSOs to be a good representation of the intrinsic QSO scatter. For the $W1W2$-bands, there could be also source confusion/deblending errors in the photometry since we just performed aperture photometry, without taking into account the large unWISE (≈6 arcsec) point spread function. A model that better reproduces the full distribution of the relative fluxes of the luminous QSOs at low-z would provide a better classification of our sources. Therefore, our conclusion is that the 'simqso' simulator was designed for colour-cuts, but it is not up to the demands of a density estimation method.

As apparent from Figs 13–15, our current simulated quasar sample fails to capture the full spectral diversity of the observed quasar population, which is important for the density estimation method. Hence, to improve on our quasar selection, we have to move beyond modeling average quasar properties, for which 'simqso' was originally designed, but rather capture the full relative flux distribution of the full population. In the future, we plan to mitigate these limitations by carefully modelling of the relative fluxes of QSOs using empirical data coming from the *SDSS* and BOSS surveys, which would capture the full distribution of quasar SEDs and hence relative fluxes.

7.2 Comparison with other probabilistic classification methods

Compared to other probabilistic classification methods, our approach has two main advantages:

(i) Our method accounts for the photometric errors by convolving the underlying density distribution with the object's uncertainties, assuming that the relative-flux uncertainties are Gaussian. While this approach is required to correctly estimate the probability that a noisy object is a member of given class, standard random forest methods ignore the photometric errors (e.g. Schindler et al. 2017; Wenzl et al. 2021), thus not utilizing all the information contained in the data. For bright sources this should not be so problematic given the small associated uncertainties. However, at high-z we have to take into account that: (1) QSOs dropout of optical bands (e.g. grz) and so we need to accurately treat low signal to noise dropout fluxes, and (2) QSOs are rare at high-z and the LFs rise with decreasing flux. So, to build-up statistics, the majority of targets will always be near the flux limits of our data, while the inclusion of the photometric errors in the analysis of fainter sources would prevent the overly optimistic identification of contaminants as high-z QSO candidates.

(ii) The method proposed by Mortlock et al. (2012) is also Bayesian, and is directly analogous to what we are doing, with the caveat that they mostly rely on constructing the models of the key contaminants (MLT dwarf types, and compact early-type galaxies). This approach requires a perfect knowledge of both the properties and the type of contaminants, whose feasibility is very challenging. For example, even if brown dwarfs and early-type galaxies are the majority among the contaminants, also Type-2 QSOs, reddened low-z QSOs, and FeLoBAL QSOs could also contaminate the high-z selection, whereas constructing models for the number density and colors of all these sources would be a daunting task. Instead, our model for the contaminant class is purely empirical and does not need to construct SED models for the mean properties of each possible contaminant.

This approach is more flexible as it captures the underlying deconvolved distribution of the contaminant using real data, and includes all the kind of possible contaminants without the need of modelling them.

8 CONCLUSION

In this paper, we described the application of the XDHZQSO method to select high-z ($6 \leq z \leq 8$) QSOs. Our approach is based on density estimation in the high-dimensional space inhabited by the optical-IR photometry. The main idea is that quasars and the far more abundant contaminants (cool dwarf stars, red galaxies, lower-z reddened, or absorbed QSOs) inhabit different regions of this space. Thus, probability density ratios yield the probability that an object is a quasar, which is used to select and prioritize candidates for spectroscopic follow-up. Density distributions are modelled as Gaussian mixtures with principled accounting of errors using the XD algorithm. Compared to other probabilistic selection methods, the great advantage of our approach is that the poorly understood contaminants are modelled fully empirically.

High-z quasars were trained on synthetic photometry in three redshift bins ($6 \leq z \leq 6.5$, $6.5 \leq z \leq 7$, $7 \leq z \leq 8$), whereas contaminants were trained on the VIKING ($YJHK_s$) imaging survey combined with deep DECaLS z-band and unWISE ($W1W2$), where all sources were required to be g and r dropouts. The combination of depth ($J_{AB} < 22$) and wide field (1076 deg^2) make this the best panchromatic imaging for training quasar selection until *Euclid* arrives.

From extensive simulations we determined the threshold ($P > P_{\rm th}$) required to obtain a completeness of $\gtrsim 56$ per cent in each redshift bin, which results in selection efficiencies $\gtrsim 5$ per cent. These high efficiencies indicate that the ≈ 1 per cent efficiencies of recent colour-cut based surveys are not necessary. The required thresholds $P_{\rm th}$ and resulting efficiencies depend on the z-bin in question owing to the changing overlap between quasars and contaminants, where the higher redshift bins have lower efficiencies. With the adopted $P_{\rm th} = 0.01$, we selected 58, 80, and 43 quasar candidates in the range $6 \leq z \leq 6.5$, $6.5 \leq z \leq 7$, $7 \leq z \leq 8$ in the VIKING footprint, respectively. These targets have been scheduled for optical and NIR spectroscopic follow-up, and the results will be published in a future work (Nanni et al. in prep.).

In the VIKING footprint the there are eight known $6 \leq z \leq 7$ QSOs that meet our catalogue criteria, of which two are selected. Since there are no $z > 7$ known QSOs in the VIKING footprint, we applied our method to six out of eight known $z > 7$ QSOs in the entire sky (we excluded two $z > 7$ QSOs as they do not meet our catalog criteria), and recover five of them. We argued that the XDHZQSO misses some of these quasars for two reasons: (1) the existing quasar fluxes are noisy so that our model correctly assigns them a low probability, and (2) the inaccuracies in our modeling of quasars, namely that the synthetic quasar spectra we used do not capture the scatter in the distribution of relative fluxes. We argued that the first limitation is a feature rather than a flaw in our approach, since we deliver reliable probabilities treating noise, and that this overall will result in higher selection efficiency. As for the second, an empirical model of luminous quasar spectra will definitely improve our classification, which we will pursue in future work.

From the integration of the $z = 6.7$ LF down to $J = 21.5$, we expect to find $\approx$15, $\approx$5, and $\approx$2 QSOs at $6 \leq z \leq 6.5$, $6.5 \leq z \leq 7$, $7 \leq z \leq 8$, respectively, in the VIKING survey area. Considering the completeness we derived in the three redshift ranges and the fact that three, and four $J \leq 21.5$ QSOs have been already discovered in the VIKING footprint at $6 \leq z \leq 6.5$, and $6.5 \leq z \leq 7$, respectively, we expect to discover $\approx$10, $\approx$1, and $\approx$2 new QSOs at $6 \leq z \leq 6.5$, $6.5 \leq z \leq 7$, $7 \leq z \leq 8$, respectively, with future spectroscopic follow-up of our candidates.

Future applications of this methodology will focus on three data sets: UKIDSS, UHS, and *Euclid*. UKIDSS covers an area of $\approx$4000 deg^2 with similar multifilter coverage as VIKING ($ZYJHK$), making it the best ground to apply XDHZQSO after VIKING. Instead, UHS covers a larger area ($\approx$12 700 deg^2) but only with three filters (JHK). To apply our method to UHS, whose sources have no data in the Y band, we will simply re-score by setting the errors in the bands with no measurements to a large number.

Finally, the advent of *Euclid* in 2022 will provide plenty of optical/IR data with a better separation between high-z QSOs and contaminants properties, as its six-yr wide survey will cover 15 000 deg^2 of extragalactic sky in four bands: a broad optical band O (5500–9000 Å), and three NIR bands, Y (9650–11920 Å), J (11920–15440 Å), and H (15440–20000 Å), a depth of 24 mag at 5σ (Laureijs et al. 2011). The *Euclid*'s wide field IR imaging should enable the discovery of $\sim$100 QSOs at $z > 7$, and $\sim$25 beyond the current record of $z = 7.6$, including $\sim$8 beyond $z = 8.0$ (Euclid Collaboration 2019). Since no data have been delivered yet from *Euclid*, we will need re-train XDHZQSO on the *Euclid* photometry to get the contaminant model. Finally, the high efficiencies in finding $z > 7$ QSOs reached by XDHZQSO suggest that we can do much more efficient spectroscopic follow-up, while we have a framework to solve the problem of performing low efficiency selection with JWST.

ACKNOWLEDGEMENTS

We thank the referee D. Mortlock for reading the paper carefully and providing useful comments. This work is part of a project that has received funding from the European Research Council (ERC) Advanced Grant program under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 885301). We thank S. Bosman and the ENIGMA group at UCSB for providing useful comments on an initial draft of this paper. We also thank J. Bovy for assistance he provided related to the XD code.

DATA AVAILABILITY

This work uses publicly available data from the VIKING DR4, the Pan-STARRS, and the DELS surveys. Links to the archives of these public surveys are provided in the main text.

REFERENCES

Aihara H. et al., 2018, PASJ, 70, S4
Bañados E. et al., 2016, ApJS, 227, 11
Bañados E. et al., 2018, Nature, 553, 473
Baldwin J. A., 1977, ApJ, 214, 679
Barnett R., Warren S. J., Cross N. J. G., Mortlock D. J., Fan X., Wang F., Hewett P. C., 2021, MNRAS, 501, 1663
Betoule M. et al., 2014, A&A, 568, A22
Bovy J., Hogg D. W., Roweis S. T., 2011a, Ann. Appl. Stat., 5, 1657
Bovy J. et al., 2011b, ApJ, 729, 141
Bovy J. et al., 2012, ApJ, 749, 41
Buitinck L. et al., 2013, ECML PKDD Workshop: Languages for Data Mining and Machine Learning. p. 108
Carnall A. C. et al., 2015, MNRAS, 451, L16
Davies F. B. et al., 2018, ApJ, 864, 142
Dey A. et al., 2019, AJ, 157, 168
Euclid Collaboration, 2019, A&A, 631, A85
Fan X. et al., 2001, AJ, 122, 2833
Gaskell C. M., 1982, ApJ, 263, 79
Glikman E., Helfand D. J., White R. L., 2006, ApJ, 640, 579
González-Fernández C. et al., 2018, MNRAS, 474, 5459
Holoien T. W. S., Marshall P. J., Wechsler R. H., 2017, AJ, 153, 249
Inayoshi K., Visbal E., Haiman Z., 2020, ARA&A, 58, 27
Jiang L. et al., 2006, AJ, 131, 2788
Jiang L. et al., 2016, ApJ, 833, 222
Kelly B. C., Bechtold J., Siemiginowska A., 2009, ApJ, 698, 895
Kuhn O., Elvis M., Bechtold J., Elston R., 2001, ApJS, 136, 225
Laureijs R. et al., 2011, preprint(arXiv:1110.3193)
Mainzer A. et al., 2011, ApJ, 743, 156
Matsuoka Y. et al., 2016, ApJ, 828, 26
Matsuoka Y. et al., 2018a, PASJ, 70, S35
Matsuoka Y. et al., 2018b, ApJS, 237, 5
Matsuoka Y. et al., 2018c, ApJ, 869, 150
Matsuoka Y. et al., 2019a, ApJ, 872, L2
Matsuoka Y. et al., 2019b, ApJ, 883, 183
McGreer I. D., Mesinger A., Fan X., 2011, MNRAS, 415, 3237
McGreer I. D. et al., 2013, ApJ, 768, 105
McGreer I. D., Mesinger A., D'Odorico V., 2015, MNRAS, 447, 499

Mortlock D. J. et al., 2011, Nature, 474, 616
Mortlock D. J., Patel M., Warren S. J., Hewett P. C., Venemans B. P., McMahon R. G., Simpson C., 2012, MNRAS, 419, 390
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Reed S. L. et al., 2015, MNRAS, 454, 3952
Reed S. L. et al., 2017, MNRAS, 468, 4702
Richards G. T. et al., 2006, AJ, 131, 2766
Richards G. T. et al., 2011, AJ, 141, 167
Ross N. P., Cross N. J. G., 2020, MNRAS, 494, 789
Schindler J.-T., Fan X., McGreer I. D., Yang Q., Wu J., Jiang L., Green R., 2017, ApJ, 851, 13
Schindler J.-T. et al., 2018, ApJ, 863, 144
Schindler J.-T. et al., 2019, ApJ, 871, 258
Schlafly E. F., Meisner A. M., Green G. M., 2019, ApJS, 240, 30
Schmidt K. B., Marshall P. J., Rix H.-W., Jester S., Hennawi J. F., Dobler G., 2010, ApJ, 714, 1194
Selsing J., Fynbo J. P. U., Christensen L., Krogager J. K., 2016, A&A, 585, A87
Vanden Berk D. E. et al., 2004, ApJ, 601, 692
Venemans B. P. et al., 2013, ApJ, 779, 24
Venemans B. P. et al., 2015, MNRAS, 453, 2259
Volonteri M., 2012, Science, 337, 544
Volonteri M., Begelman M. C., 2010, MNRAS, 409, 1022
Wang F. et al., 2017, ApJ, 839, 27
Wang F. et al., 2018, ApJ, 869, L9
Wang F. et al., 2019, ApJ, 884, 30
Wang F. et al., 2020, ApJ, 896, 23
Wang F. et al., 2021, ApJ, 907, L1
Wenzl L. et al., 2021, AJ, 162, 72
Willott C. J. et al., 2009, AJ, 137, 3541
Worseck G., Prochaska J. X., 2011, ApJ, 728, 23
Wright E. L. et al., 2010, AJ, 140, 1868
Wu X.-B. et al., 2015, Nature, 518, 512
Yang J. et al., 2016, ApJ, 829, 33
Yang J. et al., 2019, AJ, 157, 236
Yang J. et al., 2020a, ApJ, 897, L14
Yang J. et al., 2020b, ApJ, 904, 26
Yip C. W. et al., 2004, AJ, 128, 2603
Zou H. et al., 2019, ApJS, 245, 4

APPENDIX A: COVARIANCE COMPUTATION AND APPLICATION

To construct the contaminant models during the training step, we deconvolved the noisy relative fluxes of our contaminant sources, assuming that the relative-flux uncertainties are Gaussian, and providing the covariance matrix of the uncertainties of the single objects. While the flux measurements in each filter are independent of one another, i.e. their noise is uncorrelated, the relative flux errors are correlated (i.e. they are the ratio of the flux in a given band flux and the J-band flux). Thus, the covariance of a source with fluxes $\vec{f} = \{f_1, f_2, ..., f_N\}$ and uncertainties $\vec{\sigma_f} = \{\sigma_{f_1}, \sigma_{f_2}, ..., \sigma_{f_N}\}$ coming from N filters that include the J band one, can be computed as

$$
\begin{aligned}
\mathrm{cov}\left[\frac{\hat{F}_x}{\hat{F}_J}, \frac{\hat{F}_y}{\hat{F}_J}\right] &= \mathbb{E}\left[\left(\frac{\hat{F}_x}{\hat{F}_J} - \mathbb{E}\left[\frac{\hat{F}_x}{\hat{F}_J}\right]\right)\left(\frac{\hat{F}_y}{\hat{F}_J} - \mathbb{E}\left[\frac{\hat{F}_y}{\hat{F}_J}\right]\right)\right] \\
&= \mathbb{E}\left[d\left(\frac{\hat{F}_x}{\hat{F}_J}\right) d\left(\frac{\hat{F}_y}{\hat{F}_J}\right)\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\hat{F}_J}\mathrm{d}\hat{F}_x - \frac{\hat{F}_x}{\hat{F}_J^2}\mathrm{d}\hat{F}_J\right)\left(\frac{1}{\hat{F}_J}\mathrm{d}\hat{F}_y - \frac{\hat{F}_y}{\hat{F}_J^2}\mathrm{d}\hat{F}_J\right)\right] \\
&= \mathbb{E}\left[\frac{1}{\hat{F}_J^2}\mathrm{d}\hat{F}_x\mathrm{d}\hat{F}_y - \frac{\hat{F}_y}{\hat{F}_J^3}\mathrm{d}\hat{F}_x\mathrm{d}\hat{F}_J - \frac{\hat{F}_x}{\hat{F}_J^3}\mathrm{d}\hat{F}_y\mathrm{d}\hat{F}_J + \right. \\
&\quad \left. + \frac{\hat{F}_x\hat{F}_y}{\hat{F}_J^4}\mathrm{d}\hat{F}_J^2\right] \\
&= \frac{1}{\hat{F}_J^2}\mathbb{E}[\mathrm{d}\hat{F}_x\mathrm{d}\hat{F}_y] - \frac{\hat{F}_y}{\hat{F}_J^3}\left(\mathbb{E}[\mathrm{d}\hat{F}_x\mathrm{d}\hat{F}_J] + \mathbb{E}[\mathrm{d}\hat{F}_y\mathrm{d}\hat{F}_J]\right) \\
&\quad + \frac{\hat{F}_x\hat{F}_y}{\hat{F}_J^4}\mathbb{E}[\mathrm{d}\hat{F}_J^2]. \qquad \text{(A1)}
\end{aligned}
$$

In our case, the covariance matrix is:

$$
\mathrm{cov}\left[\frac{\hat{F}_x}{\hat{F}_J}, \frac{\hat{F}_y}{\hat{F}_J}\right] = \frac{\hat{F}_x\hat{F}_y}{\hat{F}_J^4}\mathbb{E}[\mathrm{d}\hat{F}_J^2] = \frac{\hat{F}_x\hat{F}_y}{\hat{F}_J^4}\sigma^2_{\hat{F}_J} \qquad \text{for} \quad x \neq y, \qquad \text{(A2)}
$$

$$
\mathrm{cov}\left[\frac{\hat{F}_x}{\hat{F}_J}, \frac{\hat{F}_y}{\hat{F}_J}\right] = \left(\frac{1}{\hat{F}_J}\right)^2\sigma^2_{\hat{F}_x} + \left(\frac{\hat{F}_x^2}{\hat{F}_J^4}\right)\sigma^2_{\hat{F}_J} \qquad \text{for} \quad x = y. \qquad \text{(A3)}
$$

At first, to train our contaminant models we provided to the XD code the noisy relative fluxes with covariance matrices computed using equations (A2) and (A3). However, we noticed that for bins whose J-band median point is $J_{\mathrm{mp}} > 21$ (i.e. $\mathrm{SNR}(J_{\mathrm{mp}}) < 10$) the XD code is not able to correctly deconvolve the contaminants properties. This is apparent in Fig. A1, where we show the comparison between the real data (black contours) and a noise added sample from the deconvolved model (red contours) generated by the XD code in a faint bin ($22.0 < J < 22.3$, $\mathrm{SNR}(J_{\mathrm{mp}}) = 5$): it is clear that we do not obtain a noisy relative flux distribution that is consistent with the real one. This deconvolved model was generated after providing a covariance matrix in the form of equation (A2) plus (A3), while we added the errors to the deconvolved sample as described in Appendix B. The failure of the XD code to correctly deconvolve the relative fluxes in the limit of faint J-band bins ($\mathrm{SNR}(J_{\mathrm{mp}}) < 10$) arises from the violation of our assumption that the relative-flux uncertainties are Gaussian in this regime. In fact, the ratio of noisy quantities is in general not Gaussian distributed, as we assumed in order to use XD. However, this is a good approximation if $\hat{F}_J$ has small errors

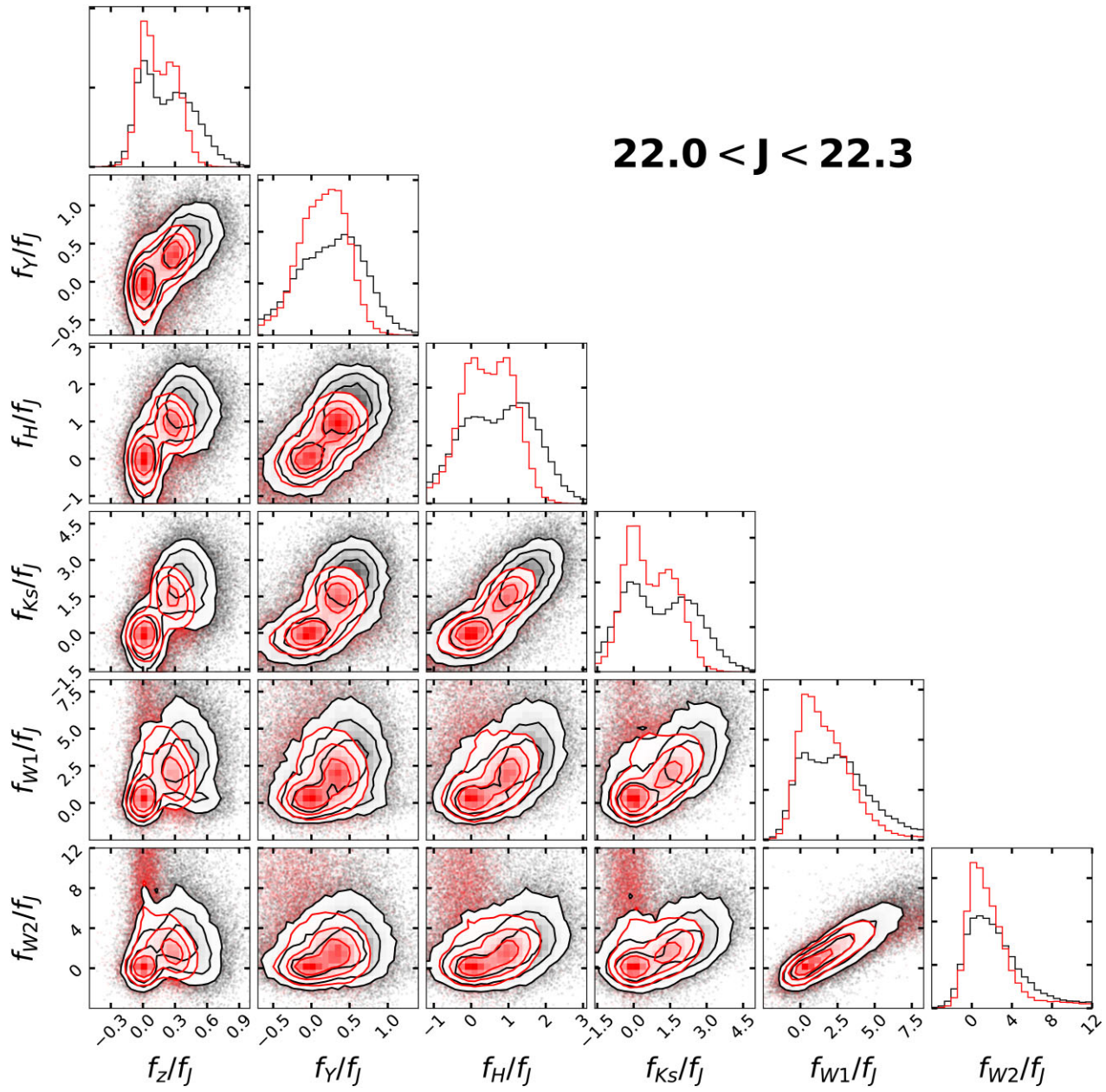

Figure A1. Relative-flux relative-flux contours comparison between the real (noisy) data (black) and a noise added sample from the deconvolved model (red) generated by the XD code in the $22.0 < J < 22.3$. Errors have been added as explained in Appendix B, while the model was generated providing a covariance matrix in the form of equation (A2) plus A3. The labelled quantities are relative fluxes (i.e. fluxes in different bands divided by the *J*-band flux). It is apparent that we do not obtain a noisy relative flux distribution that is consistent with the real one.

relative to $\hat{F}_x$, whereas as $\hat{F}_J$ becomes noisier, one will generate progressively stronger tails in $\hat{F}_x/\hat{F}_J$. To remedy this problem, we decided to construct our faint ($J_{\rm mp} > 21$) deconvolved contaminant models providing a diagonal covariance: with only elements on the diagonal computed by equation (A3) and zeros elsewhere. Although, this is not formally the correct approach to deal with non-independent quantities, it simply provides good results during the training step. In Fig. A2, we show the comparison between the real data (black contours) and a noise added sample from the deconvolved model (red contours) generated by the XD code with a diagonal covariance. The *J*-band bin and the real data are the same as those displayed in Fig. A1. In this case, it is apparent that after re-adding the errors the noisy simulated distributions are far more consistent with the real ones.

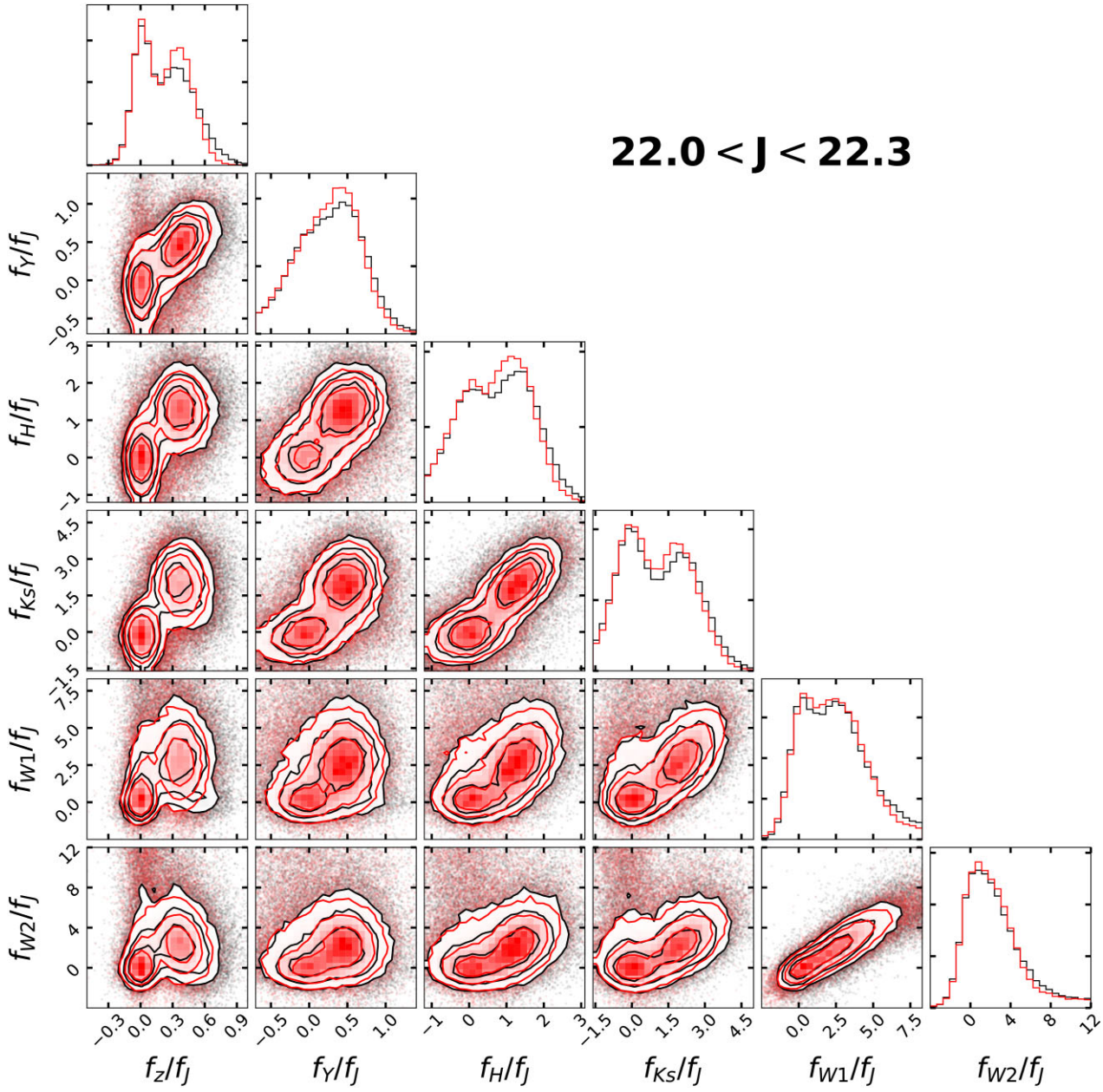

Figure A2. Relative-flux relative-flux contours comparison between the real (noisy) data (black) and a noise added sample from the deconvolved model (red) generated by the XD code in the $22.0 < J < 22.3$. Errors have been added as explained in Appendix B, while the model was generated providing a diagonal covariance matrix in the form of equation (A3). The labelled quantities are relative fluxes (i.e. fluxes in different bands divided by the *J*-band flux). In this case, the two distributions are consistent.

APPENDIX B: NOISE MODEL

As described in several parts in this paper, we often sampled a huge number of simulated high-z QSOs and contaminants from our XDHZQSO deconvolved models, and finally computed their probabilities of being high-z QSOs based on their simulated properties. However, the sampling of deconvolved models produces noiseless relative fluxes that are not a real representation of the noisy properties usually measured. We explain here our adopted procedure to add the flux uncertainties to the simulated noiseless fluxes.

Lets consider for simplicity the case of a single noiseless source sampled from our simulations of a specific *J*-band bin. The approach we describe here can then be applied to an ensemble of such samples.

For each *J*-band bin, we compute the central *J*-band flux of the bin as the median of the *J*-band fluxes of all the VIKING sources that land in the bin. Now, to generate mock photometry for the source, we multiply the median *J*-band flux for the bin with the noiseless simulated relative fluxes obtained by sampling our Gaussian mixture model, so as to obtain its noiseless fluxes in all the other bands (VIKING-YHK_s, DECaLS-z, and un*WISE*-$W1W2$). To derive the photometric error to add to these seven noiseless simulated fluxes, we start by, for each photometric band, dividing the real noisy fluxes from the VIKING data set (see description in Section 3.1) into 50 bins that roughly contain the same number of sources. For each filter and for each of these bins, we construct the cumulative distribution of the photometric error and archive them. To simulate a mock

source, we locate the bin containing its flux level for each filter, and draw samples from the respective cumulative distributions to obtain standard deviations corresponding to the noise level in each filter. We then create a realization of Gaussian noise using these standard deviations, which are then added to the noiseless mock data to construct a noisy mock observation. In this way, we can add the real errors coming from our VIKING area data set to our simulated noiseless fluxes: i.e. we capture the distribution of the noise at a given flux level, instead of simply using its mean value.

This paper has been typeset from a TEX/LATEX file prepared by the author.