The Astrophysical Journal, 747:107 (12pp), 2012 March 10 doi:10.1088/0004-637X/747/2/107


# A REFINED QSO SELECTION METHOD USING DIAGNOSTICS TESTS: 663 QSO CANDIDATES IN THE LARGE MAGELLANIC CLOUD

Dae-Won Kim[1,2,3], Pavlos Protopapas[1,3], Markos Trichas[1], Michael Rowan-Robinson[4], Roni Khardon[5], Charles Alcock[1], and Yong-Ik Byun[2,6]
[1] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
[2] Department of Astronomy, Yonsei University, Seoul, Korea
[3] Institute for Applied Computational Science, Harvard University, Cambridge, MA, USA
[4] Astrophysics Group, Imperial College, London, UK
[5] Department of Computer Science, Tufts University, Medford, MA, USA
[6] Yonsei University Observatory, Yonsei University, Seoul, Korea
*Received 2011 October 26; accepted 2011 December 31; published 2012 February 22*

## ABSTRACT

We present 663 QSO candidates in the Large Magellanic Cloud (LMC) selected using multiple diagnostics. We started with a set of 2566 QSO candidates selected using the methodology presented in our previous work based on time variability of the MACHO LMC light curves. We then obtained additional information for the candidates by crossmatching them with the *Spitzer* SAGE, the Two Micron All Sky Survey, the *Chandra*, the *XMM*, and an LMC *UBVI* catalog. Using this information, we specified six diagnostic features based on mid-IR colors, photometric redshifts using spectral energy distribution template fitting, and X-ray luminosities in order to further discriminate high-confidence QSO candidates in the absence of spectra information. We then trained a one-class Support Vector Machine model using the diagnostics features of the confirmed 58 MACHO QSOs. We applied the trained model to the original candidates and finally selected 663 high-confidence QSO candidates. Furthermore, we crossmatched these 663 QSO candidates with the newly confirmed 151 QSOs and 275 non-QSOs in the LMC fields. On the basis of the counterpart analysis, we found that the false positive rate is less than 1%.

*Key words:* Magellanic Clouds – methods: data analysis – quasars: general

*Online-only material:* machine-readable table

## 1. INTRODUCTION

Active galactic nuclei (AGNs) are very energetic extragalactic objects that have been studied in many astronomical fields such as galaxy formation and evolution (e.g., Heckman et al. 2004; Bower et al. 2006; Trichas et al. 2009, 2010), large-scale structure (e.g., Ross et al. 2009), dark matter substructure (e.g., Miranda & Macciò 2007), and black hole growth (e.g., Kollmeier et al. 2006).

It is known that QSOs show strong variability over a wide range of wavelengths on a timescale from a few days to several years (Hook et al. 1994; Hawkins 2002). It is widely believed that the variability is associated with accretion disk instability (Rees 1984; Kawaguchi et al. 1998). Recently, interesting studies on QSO variability have been published (Kelly et al. 2009; MacLeod et al. 2010), which confirmed a correlation between the timescale of QSO variability and the physical parameters of QSOs such as black hole mass. Although these studies confirmed the correlation, different studies showed a discrepancy at the timescales of QSO variability (Kelly et al. 2009; Kozłowski et al. 2010; MacLeod et al. 2010). Possible reasons for the discrepancy are (1) poorly sampled light curves and/or short observational periods, (2) false positives such as stellar contaminations in their QSO candidates, and (3) biased QSO samples in luminosity or black hole mass. Thus, having a well-sampled set of QSO light curves with a long baseline and small number of false positives is critical for the comprehensive analysis of this correlation. Note that there are only a few hundred well-sampled QSO light curves, and a large portion of them are around the Large Magellanic Cloud (LMC) fields where the MACHO survey monitored for several years (e.g., see Geha et al. 2003; Kelly et al. 2009; Kozłowski et al. 2012).

The MACHO survey observed the sky around the LMC for 7.4 years with relatively regular sampling of a few days. The majority of the MACHO light curves have more than several hundred data points and therefore the MACHO light curves are suitable for the QSO variability studies. Nevertheless, there are only 59 confirmed MACHO QSOs in the 40 $\deg^2$ areas around the LMC (Geha et al. 2003). The main reasons for the relatively small number of QSOs are (1) the crowdedness of the fields, which makes it difficult to select QSO candidates among the dense stellar sources and thus yields a high false positive rate (e.g., see Geha et al. 2003; Dobrzycki et al. 2005), and (2) the high cost of spectroscopic or X-ray observations, which are the best methods for confirming QSOs. Thus a novel QSO selection algorithm with a high efficiency and a low false positive rate is essential to make the best use of the expensive spectroscopic telescope time and increase the collection of QSOs.

In our previous work (Kim et al. 2011), we developed a QSO selection method using a supervised classification model trained on a set of variability features extracted from the MACHO light curves including a variety of variable stars, non-variable stars, and QSOs. The trained model showed a high efficiency of 80% and a low false positive rate of 25%. Using this method, we first selected 2566 QSO candidates from the light curve database. We then developed and employed a decision procedure on the basis of diagnostics using (1) mid-IR colors, (2) photometric redshifts, and (3) X-ray luminosities on these candidates in order to separate *high-confidence* QSO candidates (hereinafter hc-QSOs). As a result, we chose in total 663 hc-QSOs out of 2566. These 663 candidates are likely QSOs; if confirmed this will increase the previous collection of QSOs in the MACHO LMC database by a factor of ∼12. Note that most of the hc-QSO light curves are well sampled for 7.4 years (i.e., several

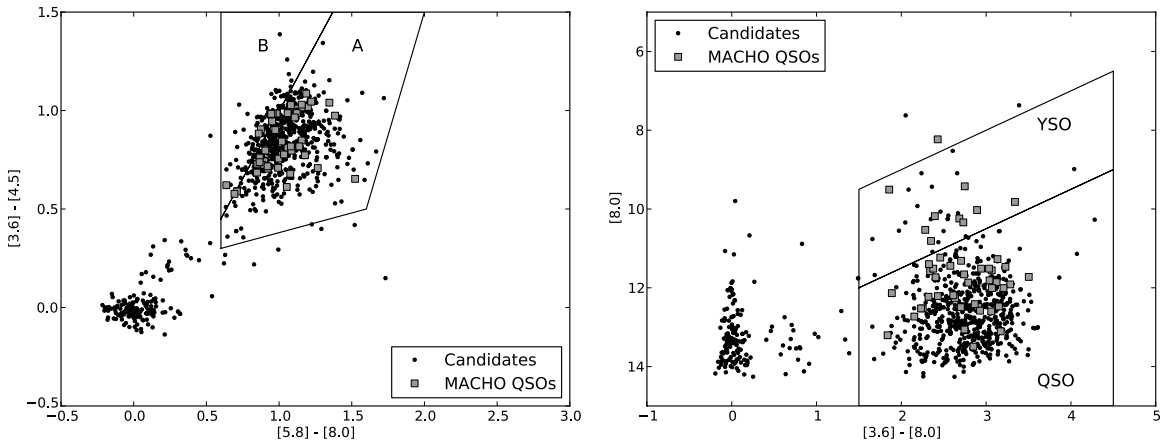


**Figure 1.** Mid-IR color–color and color–magnitude diagrams of the *Spitzer* SAGE counterparts with our QSO candidates (dots). Each axis of the figure is either *Spitzer* magnitude or color. All sources inside the four regions A, B, QSO, and YSO are potential QSOs (Kozłowski & Kochanek 2009). There are 469 candidates inside both the QSO and A regions, which are the most promising QSO candidates. The confirmed MACHO QSOs are also inside these four regions (boxes).

hundred data points with relatively regular sampling) and are chosen in such a way to exclude any potential false positives. Therefore, the light curve collection of hc-QSOs is a valuable set for QSO variability studies and can be used as a target set for spectroscopic observations.

In Section 2, we briefly introduce the MACHO database and the QSO selection algorithm that we developed to select the initial set of QSO candidates. We then present multiple diagnostics that we applied on the set of QSO candidates in Section 3. Section 4 presents a classification model trained on the diagnostics features in order to choose hc-QSOs. In Section 5, we crossmatch our candidates with newly discovered QSOs in the LMC fields. A summary is given in Section 6.

## 2. QSO CANDIDATES IN THE MACHO LMC DATABASE

We first selected QSO candidates from the MACHO light curve database using the QSO selection method developed by Kim et al. (2011; hereinafter K-method). In this paper, we used a 10% QSO probability product cut to select the QSO candidates rather than a 25% cut which Kim et al. (2011) used because we will employ other diagnostics (see Section 3) that are able to effectively remove false positives.[7] Here probability product is the product of the probabilities derived independently from MACHO *B*- and *R*-band light curves using Support Vector Machine (SVM; Boser et al. 1992) and Platt's probability estimation (Platt 1999). By definition, QSO candidates with higher probabilities are more likely to be QSOs. With the probability cut of 10%, we found 2566 QSO candidates.

## 3. DIAGNOSTICS OF THE QSO CANDIDATES

In the following subsections, we will introduce the diagnostics performed and the consequent results.

### 3.1. Spitzer Mid-IR Properties

It is known that mid-IR color selection is an efficient discriminator for AGNs and stars/galaxies resulting from the fact that the spectral energy distributions (SEDs) of these sources are substantially different from each other (Laurent et al. 2000; Lacy et al. 2004). Lacy et al. (2004) introduced a mid-IR color cut to separate AGNs using *Spitzer* SAGE (Surveying the Agents of a Galaxy's Evolution; Meixner et al. 2006) catalog. Kozłowski & Kochanek (2009) employed a similar mid-IR color cut and selected about 5000 AGN candidates from the *Spitzer* SAGE catalog.

We used these mid-IR color selections as the first diagnostic. We crossmatched our candidates with the *Spitzer* SAGE LMC catalog containing 6 million mid-IR objects in order to check whether our candidates are inside the mid-IR selection cuts. We searched for the nearest SAGE source from each candidate within a 1″ search radius. In order to minimize false crossmatchings, we defined a source as a counterpart only if there are no other *Spitzer* sources within a 3″ radius from the candidate.

We found about 700 *Spitzer* counterparts shown in Figure 1 (dots). The sources inside region B could be either AGNs or stars, while the sources inside region A are likely AGNs. The YSO region is thought to be dominated by young stellar objects (YSOs) while the QSO region is thought to be dominated by AGNs. Nevertheless, all the sources inside these four regions are potential QSOs.[8] Almost all of the confirmed MACHO QSOs are inside these four regions as shown in Figure 1 (boxes).[9] The candidates inside these regions are most likely broad emission line QSOs (i.e., Type I AGNs; Stern et al. 2005). Among these counterparts, the sources inside both the QSO and the A regions are likely to be QSOs. We found that 469 QSO candidates are inside both QSO and A regions.

Figure 2 shows the estimated K-method QSO probability products of these 469 candidates. As the histogram shows, there are more QSO candidates at higher probability than lower probability, which implies that the mid-IR diagnostic is in line with the K-method.[10] In addition, the histogram shows a bimodal distribution of the probabilities. We will address this bimodality in the following section.

### 3.2. Photometric Redshift Using Template Fitting

We first crossmatched the 2566 QSO candidates with the *UBVI* catalog for the LMC (Zaritsky et al. 2004) and the Two Micron All Sky Survey (2MASS) catalog (Skrutskie et al.

[7] A lower probability cut typically produces not only more QSO candidates but also more false positives.

[8] The strongest statement is that QSOs are very unlikely to be outside those four regions.

[9] There are 48 MACHO QSOs that were crossmatched with the SAGE catalog.

[10] In the case of the entire 2566 QSO candidates, the number of candidates decreases at higher probability.

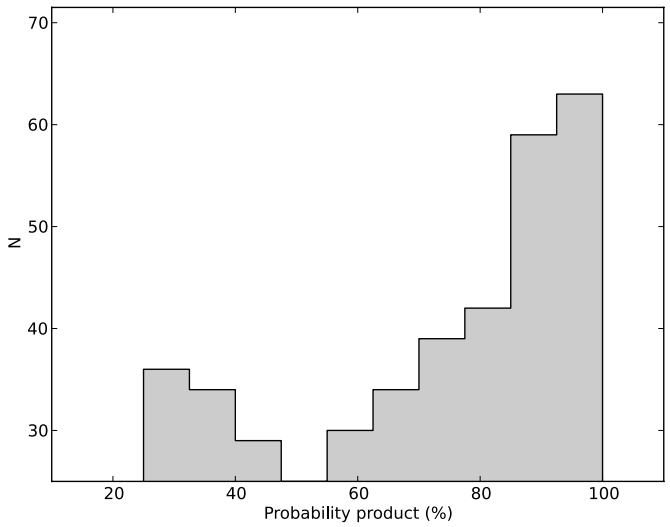


**Figure 2.** Histogram of K-method QSO probabilities of the SAGE counterparts inside both the QSO and the A (see Figure 1). There are more high-probability candidates than low-probability candidates, which indicates that the candidates inside the QSO and the A are likely to be QSOs. The histogram also shows a bimodal distribution as is addressed in Section 3.2.

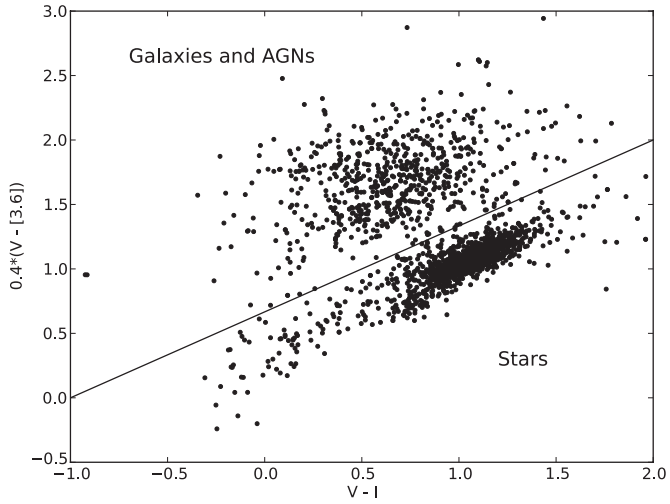


**Figure 3.** Criterion (the solid line) to separate extragalactic sources ("galaxies and AGNs" in the figure) from stars (Eisenhardt et al. 2004; Rowan-Robinson et al. 2005). Using the criterion, 686 candidates were classified as extragalactic sources (above the line) and 1274 candidates were classified as stars (below the line).

2006) to extract *UBVI* and *JHK* magnitudes. We searched the nearest source from each of the candidates within a 3″ search radius. In the case of the *UBVI* catalog, we found in total 2375 counterparts. Among them, 84% (93%) *UBVI* counterparts are within a 1″ (1.″5) distance from the candidates. In addition, only 0.3% (2% or 17%) of the candidates have another counterpart within a 1″ (1.″5 or 3″) distance from the candidates. Thus the portion of the false crossmatching is not significant. In the case of the 2MASS catalog, we found in total 846 counterparts. From those, 74% (83%) are within a 1″ (1.″5) distance from the candidates while 0% (0.1% or 0.5%) of the candidates have another counterpart within a 1″ (1.″5 or 3″) distance from the candidates. Again the portion of the false crossmatching is negligible.

We then separated stars from "galaxies and AGNs" (i.e., extragalactic sources) using a criterion proposed by Eisenhardt et al. (2004) and Rowan-Robinson et al. (2005). Figure 3 shows the criterion (the solid line) we applied. There were 686 extragalactic sources (above the cut) and 1274 stars (below the cut).[11] These 686 extragalactic sources were then fitted with galaxy templates in order to derive photometric redshifts (Rowan-Robinson et al. 2008). The templates contained three QSOs, one starburst, and ten galaxy templates. For details about the photometric redshift estimations and the SED template fitting, see Rowan-Robinson et al. (2008).

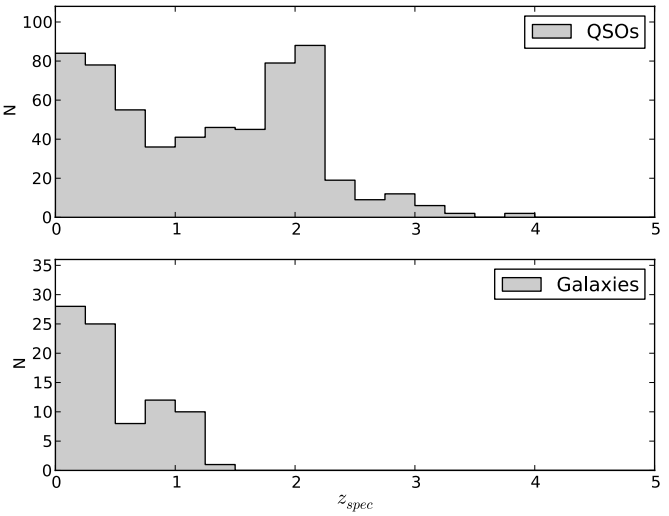


**Figure 4.** Photometric redshifts of the 602 QSO candidates fitted with the AGN templates (top panel) and the 84 QSO candidates fitted with the galaxy templates (bottom panel; Rowan-Robinson et al. 2008). The 602 QSO candidates show relatively larger redshifts than the 84 candidates.

Among the extragalactic sources, 602 were fitted with AGN templates (i.e., QSOs) while the remaining 84 were fitted with the galaxy templates (i.e., galaxies). These 602 candidates are likely QSOs. Figure 4 shows the photometric redshifts of these QSOs and galaxies. As the figure shows, the QSOs (top panel) have relatively higher redshifts than the galaxies (bottom panel). QSOs are much more luminous than galaxies and thus are detectable at higher redshifts than galaxies. In Figure 5, we show the comparison between the photometric redshifts and the spectroscopic redshifts of the confirmed MACHO QSOs (Geha et al. 2003). Out of the 58 confirmed MACHO QSOs,[12] 40 are fitted with the photometric redshift code. The remaining 18 were not fitted due to the lack of data (i.e., *UBVI* magnitudes). Among these 40 confirmed MACHO QSOs, only one was best fitted with galaxy templates while the other 39 were fitted with AGN templates. The QSO best fitted with the galaxy templates is confirmed to be a QSO from the work done by Schmidtke et al. (1999) and Geha et al. (2003). Out of the 40 QSOs, 28 (70%) are inside the ±0.1 dex accuracy (the dashed line in the figure).

Figure 6 shows the K-method probability of QSOs, galaxies, and stars discriminated during the photometric redshift estimation. As the figure shows, the majority of QSOs have higher probabilities than galaxies and stars, which implies that galaxies and stars have different and most likely weaker variability characteristics from/than QSOs. Note that the probabilities are from the K-method which mainly used variability features of light curves to select QSO candidates.

[11] We excluded the sources that do not have enough color information.

[12] Note that 58 of 59 MACHO QSOs had been monitored more than several hundred times during 7.4 years of observation while the remaining one MACHO QSO has only about 50 data points. We excluded the QSO with 50 data points from the analysis in this paper.

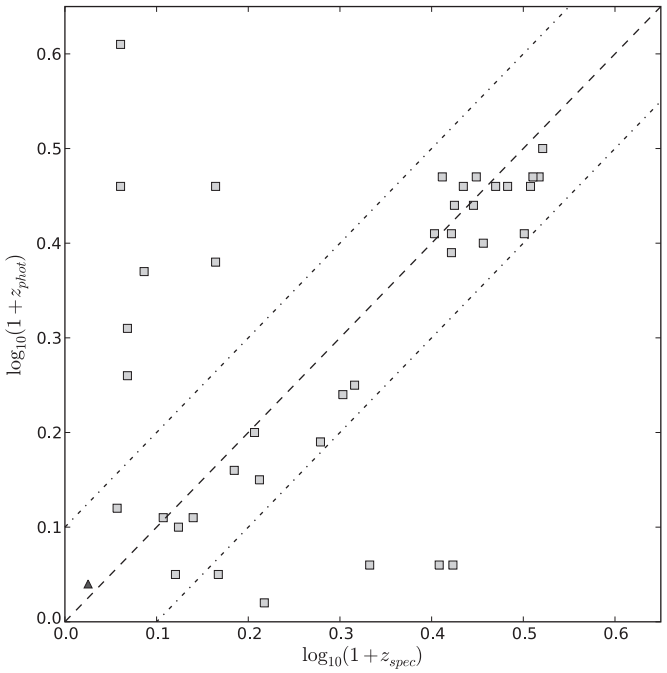


**Figure 5.** Comparison between the spectroscopic redshifts (Geha et al. 2003) and the photometric redshifts for the confirmed MACHO QSOs. Seventy percent of estimated redshifts are well matched with the spectroscopic redshifts (see the dashed line corresponding to ±0.1 dex accuracy). There is one MACHO QSO (triangle) that is fitted with the galaxy templates and 39 MACHO QSOs (squares) that are fitted with the AGN templates (Rowan-Robinson et al. 2008).

The left panel of Figure 6 also shows similar bimodality as seen in Figure 2. In order to check if there exists (1) different variability characteristics between QSOs, galaxies, and stars, and (2) different variability characteristics between the high- and low-probability QSO candidates, we show histograms of two variability features defined in Kim et al. (2011) in Figure 7. The left 2 × 2 sub-panels (left panels A, B, C, and D) show the histogram of $\sigma/\bar{m}$, where $\sigma$ is the standard deviation and $\bar{m}$ is the mean magnitude. In general, $\sigma/\bar{m}$ is large when a light curve has strong variability. The $x$-axis is scaled to be between 0 and 1. To check if differences exist between high- and low-probability QSOs (A and B), we selected two subsets: one of high ($\geqslant$80%) and the other of low ($\leqslant$40%) probability QSOs. We included all galaxies (C) and stars (D) regardless of their probabilities. As the left panels display, galaxies and stars show different distributions from the distribution of QSOs that has a peak around ∼0.3. Nevertheless, high- and low-probability QSOs do not show different distribution. The right 2 × 2 sub-panels (right panels A, B, C, and D) show a different time variability index, Stetson $K_{AC}$, which is the observation of the distribution of data points between the maximum and minimum values of the autocorrelation function of a light curve (Kim et al. 2011). As the panels show, high-probability QSOs (A) show a peak around 0.6 while low-probability QSOs (B) show a peak around 0.4. Galaxies (C) and stars (D) show peaks around 0.7. Thus, it seems that the bimodality shown in the left panel of Figure 6 and the different distributions between QSOs, galaxies, and stars in Figure 6 is correlated with the different variability characteristics of the light curves. Further analysis of this bimodality, requiring careful investigation of many variability characteristics and understanding of the selection biases, is beyond the scope of this paper.

In addition, Figure 8 shows the mid-IR colors of QSOs, galaxies, and stars. As the figure shows, almost all of the QSOs (dots) are inside the four regions while most of the stars (triangles) are outside the regions. Galaxies (squares) are either inside or outside the regions.

### 3.3. X-Ray Luminosity

In order to estimate the X-ray luminosity, we crossmatched the 2566 QSO candidates with two X-ray point-source catalogs: the *Chandra* X-ray source catalog (Evans et al. 2010) and the *XMM-Newton* 2nd Incremental Source catalog (Watson et al. 2009). We searched for the nearest source within a 5″ search radius from each candidate. The majority of the crossmatched counterparts were within a 3″ distance from the candidates and there were no additional counterparts within a 5″ distance from the candidates. We found 88 counterparts from either the *XMM* or *Chandra* catalogs.

Among the 88 counterparts, 64 were fitted with the SED templates mentioned in Section 3.2 and therefore had estimated photometric redshifts. We used the photometric redshifts and X-ray fluxes from the catalogs to calculate the X-ray luminosity of each counterpart. Figure 9 shows the photometric redshifts ($x$-axis) and the estimated X-ray luminosity, $\log L_X$ ($y$-axis).

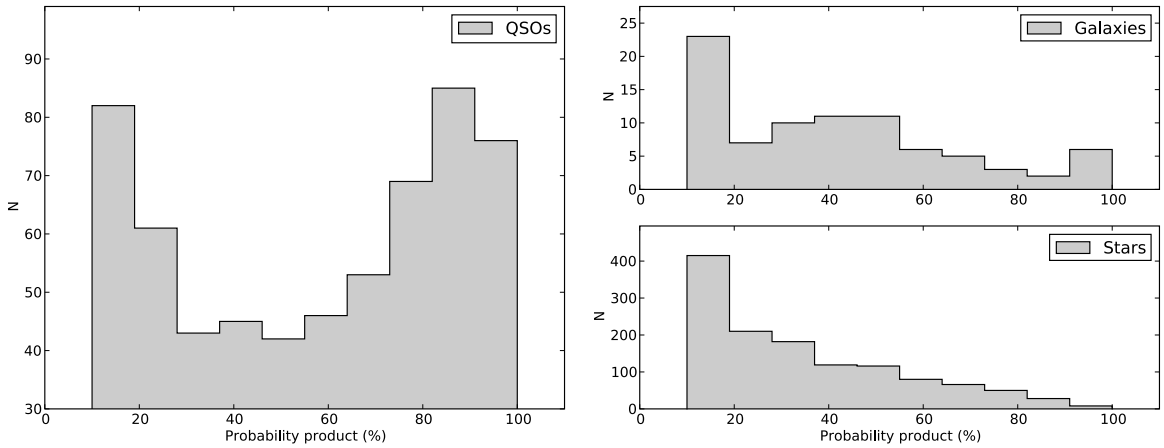


**Figure 6.** Left: histogram of the estimated K-method QSO probabilities for 602 QSOs fitted with the AGN templates. The histogram shows a bimodal distribution similar to the histogram shown in Figure 2. The bimodality is correlated with different variability characteristics of the low- and high-probability QSO candidates. See the text and Figure 7 for details. Right: histogram of the estimated K-method QSO probability of 84 galaxies (top panel) and 1274 stars (bottom panel) separated using a approach proposed by Eisenhardt et al. (2004) and Rowan-Robinson et al. (2005). As the histogram clearly shows, they have relatively lower probabilities than QSOs.

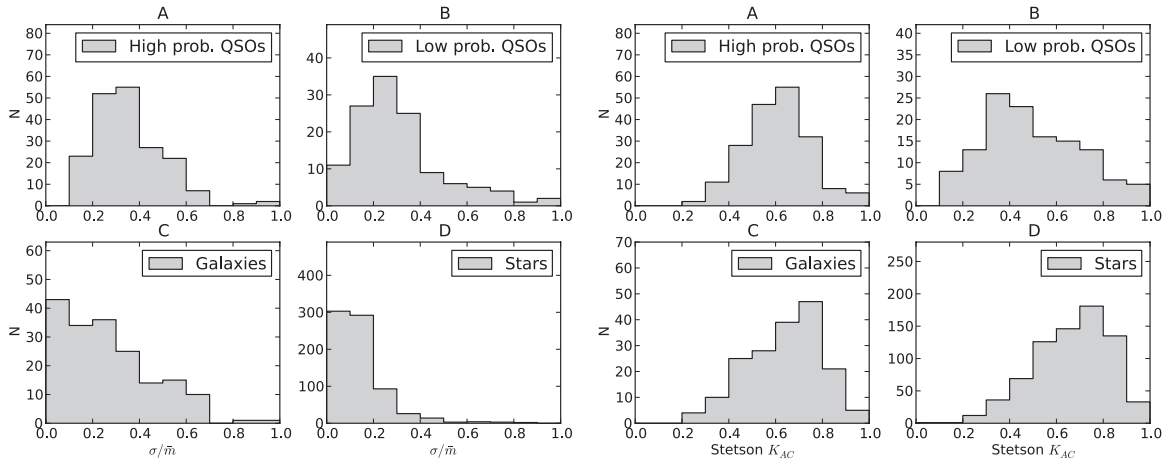


**Figure 7.** Left panels A, B, C, and D: histogram of one of the time series features, $\bar{\sigma}/\bar{m}$ (Kim et al. 2011). Galaxies and stars show different distribution from both high- and low-probability QSOs while high- and low-probability QSOs do not show distinctive differences. Right panels A, B, C, and D: histogram of Stetson $K_{AC}$ (Kim et al. 2011). High-probability QSOs show different distribution from low-probability QSOs while galaxies and stars show almost identical distributions. As the histograms show, it seems that the bimodality in the left panel of Figure 6 is correlated with the different variability characteristics of each class. Further analysis of this bimodality is beyond the scope of this paper.

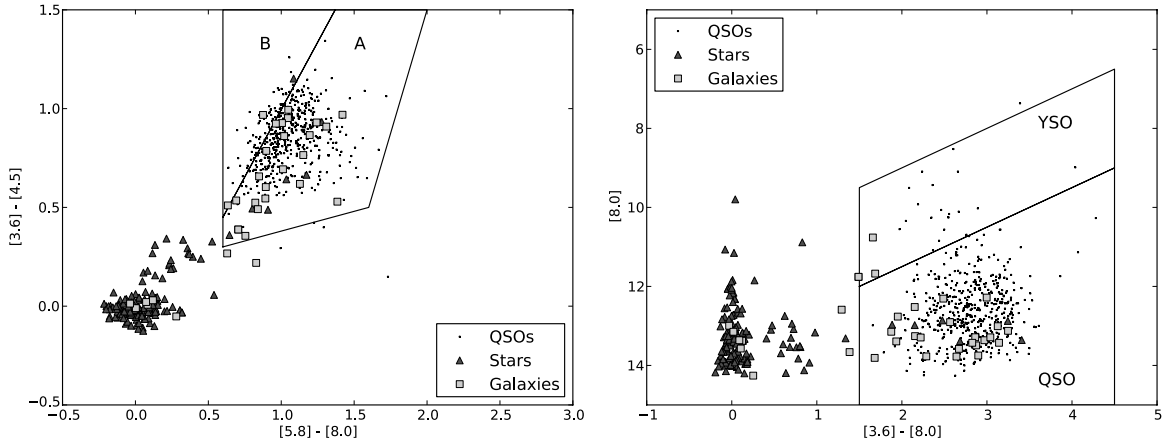


**Figure 8.** Mid-IR color–color and color–magnitude diagrams of the QSOs, galaxies, and stars classified using the photometric redshift code. See Section 3.2 for details. Each axis of the figure is either *Spitzer* magnitude or color. In the left panel, there are 502 QSOs (dots), 33 galaxies (squares), and 145 stars (triangles). In the right panel, there are 518 QSOs, 34 galaxies, and 145 stars. As the panels show, almost all of the QSOs and galaxies are inside the regions (QSO, YSO, A, and B), which indicates that all of them are potential QSOs. On the other hand, the majority of the stars are outside the regions.

In the left panel, we show the 61 *XMM* counterparts including eight confirmed MACHO QSOs. The right panel shows 14 *Chandra* counterparts including three confirmed MACHO QSOs. Almost all of the candidates (60) have higher $\log L_X$ than 42. In addition, six confirmed MACHO QSOs and 26 candidates show $\log L_X$ higher than 44. The candidates showing higher $\log L_X$ than 44 (42) are likely to be QSOs (AGNs; Elvis et al. 1994; Persic et al. 2004). The remaining candidates that show lower $\log L_X$ than 42 are likely to be galaxies.

We show the mid-IR colors of these X-ray counterparts in Figure 10. The classification of QSOs (dots), AGNs ("×"s), and galaxies (squares) is based on the X-ray luminosity of the counterparts.

## 4. HIGH-CONFIDENCE QSO CANDIDATE SELECTION USING SUPPORT VECTOR MACHINES

### 4.1. Support Vector Machine

SVM (Boser et al. 1992) is a supervised machine learning algorithm that trains a two-class classification model using samples of two known classes (i.e., training set). SVM is currently one of the best classification methods in machine learning. The classifier of an SVM defines a linear hyperplane that separates two classes in training data. To select a unique hyperplane among the set of possible hyperplanes that separate the data, SVM chooses the hyperplane which maximizes the margin between the two classes, and is therefore often called the *maximum margin separator*. SVM is also able to separate nonlinearly separable classes by using a kernel function (e.g., a polynomial kernel or a radial basis kernel), transforming nonlinear feature spaces into linear feature spaces. The hypothesis of SVM has the form

$$\mathrm{Class}(z) = \mathrm{sign}\left(\sum_i \alpha_i y_i K(z, x_i) - b\right), \quad (1)$$

where $i$ are the indices for training set examples, $x_i$ are the examples, $y_i$ are the labels, $z$ is the example that we are predicting the label for, $K(z, x_i)$ is a kernel function, and $b$ is a threshold. The $\alpha_i$ are the parameters learned by the training procedure. Despite the mapping to a potentially high dimensional space using a kernel function, the maximum margin criterion leads to automatic capacity control and thus avoids overfitting.

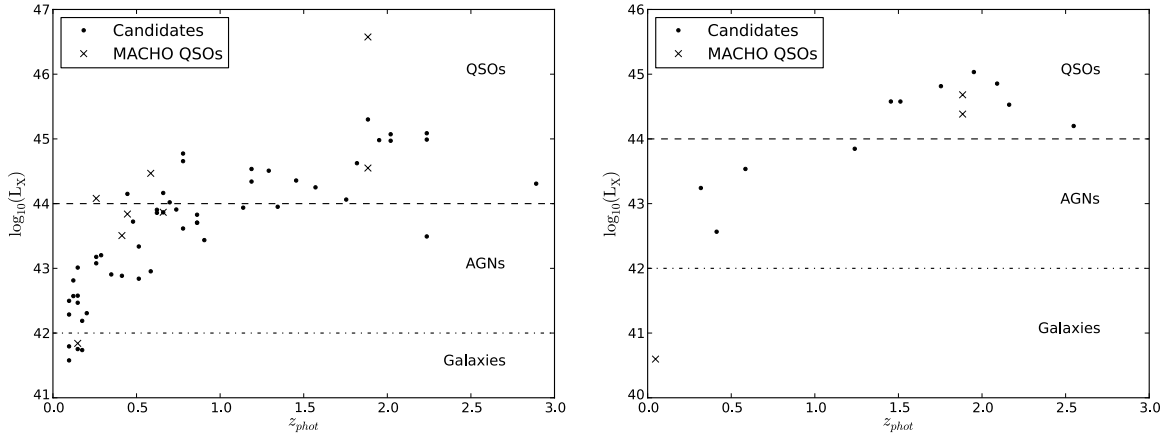


**Figure 9.** Scatter plot of the photometric redshifts ($x$-axis) and the estimated X-ray luminosity, $\log L_X$, ($y$-axis). The dots are our QSO candidates and the "×"s are the confirmed MACHO QSOs. Left: *XMM* counterparts. Right: *Chandra* counterparts. As the figures show, most of our candidates and MACHO QSOs have $\log L_X > 42$, which indicates that they are likely QSOs.

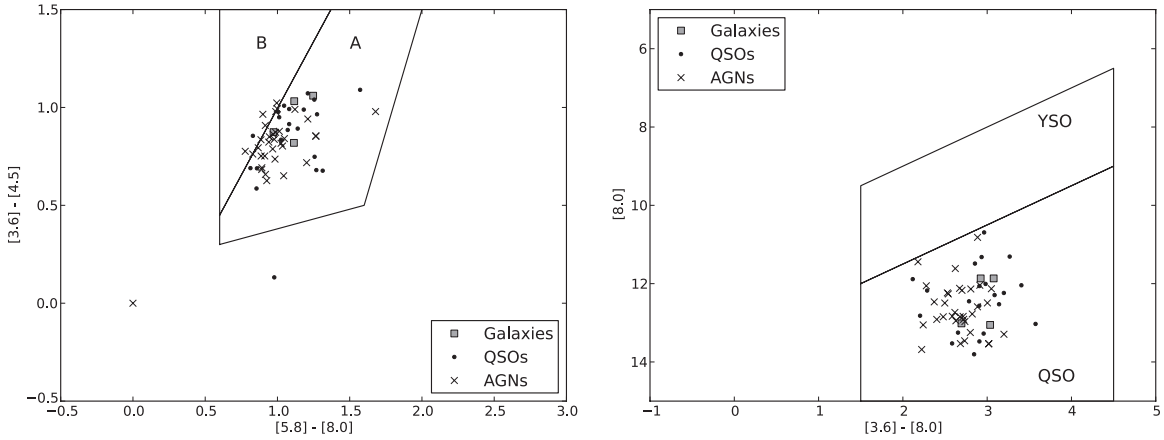


**Figure 10.** Mid-IR color–color and color–magnitude diagrams of the QSOs, galaxies, and stars classified using the X-ray luminosity. See Section 3.3 for details. Each axis of the figure is either *Spitzer* magnitude or color. As the panels show, almost of the X-ray counterparts are within the QSO and the A region. The candidates inside the QSO and the A region are very likely QSOs (Kozłowski & Kochanek 2009).

Compared to neural networks, SVMs provide a flexible classification model, avoid the problems of local minima, and reduce the need for parameter tuning. For an overview, discussion, and practical details, see Cristianini & Shawe-Taylor (2000), Bennett & Campbell (2000), Hsu et al. (2003), Kim et al. (2011), and references therein. Because the standard SVM can only solve a two-class problem, Schölkopf et al. (2001) proposed a method to solve one-class classification problems using the SVM. In brief, they define the origin as the second class and separate the one class from the origin using the SVM. For details about the method, see Schölkopf et al. (2001) and Manevitz & Yousef (2002).

### *4.2. Training a One-class SVM to Select High-confidence QSO Candidates*

We employed the one-class SVM classification method to select high-confidence QSO candidates because we do not have negative examples (i.e., non-QSO training set). We used a linear kernel rather than a polynomial kernel or a radial basis kernel because we empirically found that using other kernels did not improve classification results. To train a model, we first defined the diagnostics results as feature vectors. Table 1 summarizes the feature vectors. When we could not determine a feature value due to the nonexistence of counterpart with either the *Spitzer* SAGE, *UBVI*, or X-ray catalogs, we assigned zero to the corresponding feature. Figure 11 outlines the calculation of the diagnostics and the number of candidates for which the diagnostics are available. As mentioned above, we started with the 2566 QSO candidates selected using the K-method ("Data Preparation" panel in the figure). The diagnostics applied to these candidates are shown in the "High Confidence QSO Selection" panel. We also show the number of QSO candidates after the diagnostics (double-lined rectangles).

We trained a one-class SVM model using these features.[13] We then tuned the model by adjusting the threshold, $b$, in order to (1) obtain the highest efficiency based on the confirmed 58 MACHO QSOs, and (2) minimize the number of selected QSO candidates, which reduces the number of false positives as well. Figure 12 shows the efficiency and the number of candidates as a function of $b$. The black square shows the threshold we finally adopted. Using the determined threshold, the trained model showed 74% efficiency. We applied the tuned model to the 2566 QSO candidates and selected 663 QSO candidates (i.e., hc-QSOs).

Table 2 shows a few important parameters for some of the QSO candidates. The entire parameters of the 2566 QSO candidates are published in the online Journal. We also

[13] We used the LIBSVM package (Chang & Lin 2001).

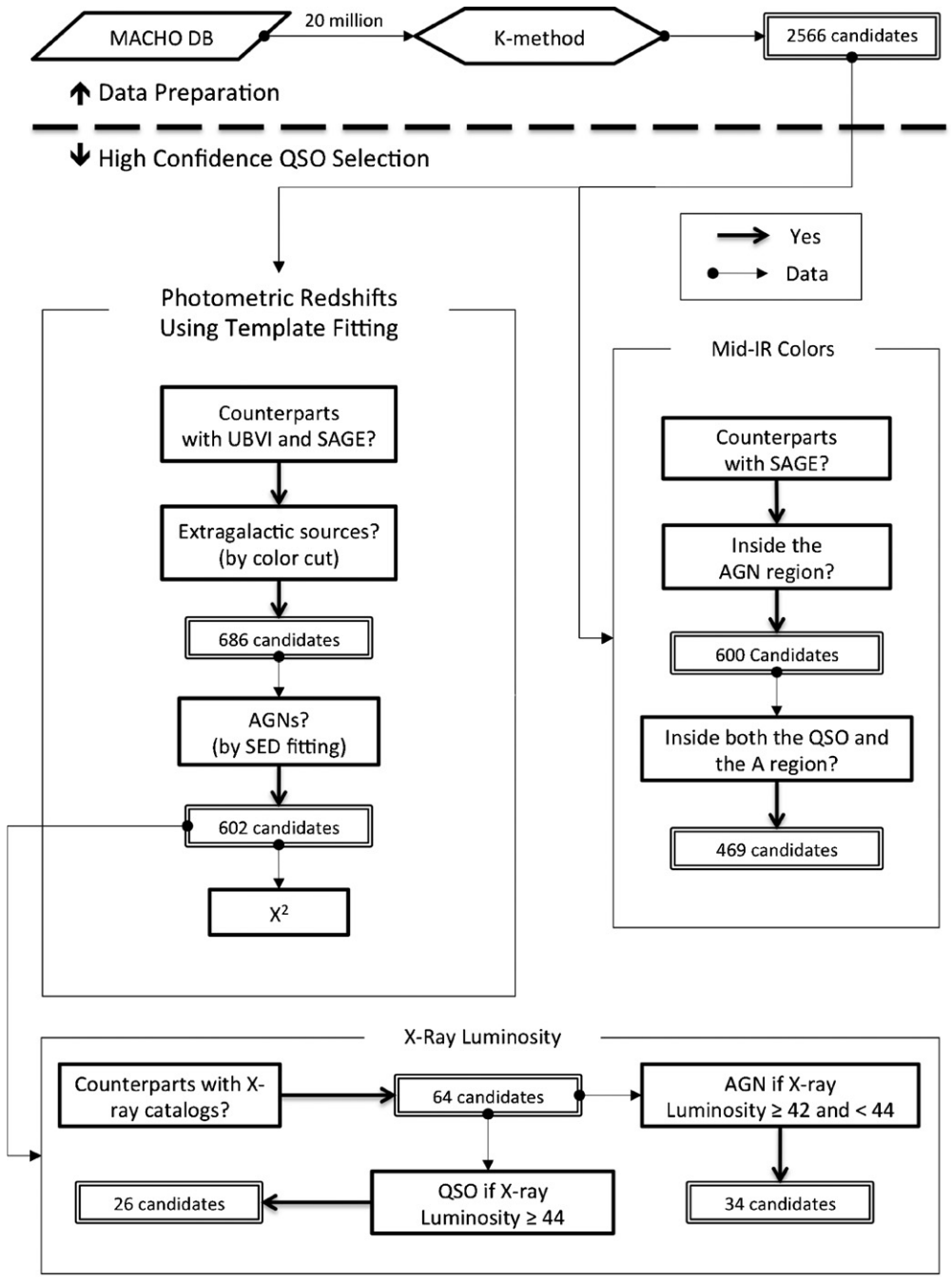


**Figure 11.** Illustration of the processes that we used to select hc-QSOs. The rectangles with bold borderlines are the diagnostics. At most of the diagnostics, we determined if the candidates are likely to be QSOs (solid line arrows). The thin arrows show the data flow. The double-lined rectangles show the number of candidates.

**Table 1**
Feature Vectors

| Mid-IR | Extragalactic Sources/Stars | SED Fitting | $\chi^2$ | *Chandra* | *XMM* |
|---|---|---|---|---|---|
| No CP[a]: 0 | No CP: 0 | No CP: 0 | No CP: 0 | No CP: 0 | No CP: 0 |
| Inside any of the four regions: 1 | Stars: 1 | Galaxies: 1 | $\chi^2$ value[b] | Galaxies: 1 | Galaxies: 1 |
| Inside both the QSO and A region: 2 | Extragalactic sources: 2 | AGNs: 2 | | AGNs: 2 | AGNs: 2 |
| | | | | QSOs: 3 | QSOs: 3 |

**Notes.**
[a] No counterpart.
[b] $\chi^2$ is from the SED fitting.

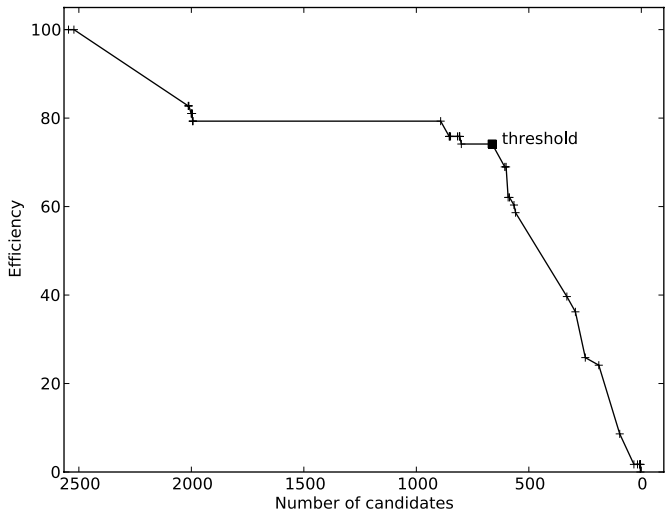


**Figure 12.** Efficiency vs. number of selected QSO candidates as a function of the SVM threshold, $b$. The black square shows the final threshold we adopted.

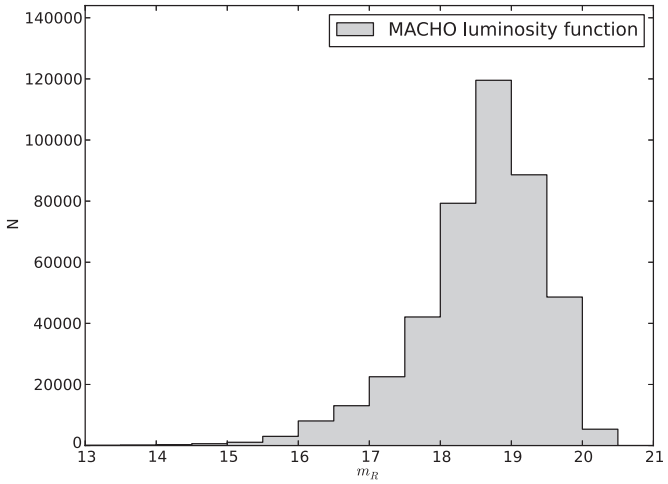


**Figure 13.** Luminosity function of MACHO $R$ magnitude from one MACHO field. The $x$-axis is MACHO $R$ magnitude and the $y$-axis is the number of MACHO sources. As the figure shows, the limiting $R$ magnitude is around 19–19.5.

provide catalogs and light curves of all the candidates at http://timemachine.iic.harvard.edu/coati/QSOs.

## 5. CROSSMATCHING WITH NEWLY DISCOVERED QSOs BY KOZŁOWSKI (2012)

Recently, Kozłowski et al. (2012) selected QSO candidates using mid-IR colors, X-ray emission, and/or optical variability in the OGLE light curve database (Udalski et al. 2008). For the variability selection, they used the Damped Random Walk (DRW) model of light curves (Kelly et al. 2009; Kozłowski et al. 2010) and then applied several cuts including magnitude, model fitting accuracy,[14] slope of a structure function, amplitude, and timescale of light curve variations. They then visually examined all the light curves of the candidates and removed about 96% of light curves (∼23,000) from the final list. Most of false positives were the "ghost" variable objects caused by photometric defects. They finally observed 845 QSO candidates using AAT/AAOmega[15] and confirmed 169 QSOs including 25 previously known QSOs[16] (i.e., 144 newly discovered QSOs) in the four ∼3 deg$^2$ fields near the LMC center. They also provided the list of remaining 676 objects. Among these 676 objects, they confirmed that 275 are non-QSOs, including YSOs, red stars, blue stars, Be stars, and planetary nebulae.[17]

To estimate the efficiency and the false positive rate of our selection method, we first crossmatched the 151 discovered QSOs[18] and 275 confirmed non-QSOs (i.e., false positives) with the entire MACHO LMC light curve database. We searched the nearest MACHO LMC source within a 3″ search radius. Out of 151 QSOs and 275 non-QSOs, 64 and 122 were crossmatched with the MACHO sources, respectively. Note that only 46 out of 64 were selected using variability characteristics in the OGLE-III light curves (Kozłowski et al. 2012).

Among these 46 QSOs, 20 are in the hc-QSO list (hereinafter c-QSOs) and 26 are not in the hc-QSO list (hereinafter cn-QSOs), which gives us 43% efficiency. It is worth mentioning that the yield of QSO candidates from Kozłowski et al. (2012) selected using only variability based on the DRW model was 7%.

Despite the fact that these 46 QSOs were determined to be variable objects based on the optical OGLE-III light curves, some of them do not show strong variability in the MACHO light curves because of (1) the difference of the limiting magnitudes of the two surveys, and (2) the photometric uncertainty of the MACHO light curves. For instance, we found that 11 of cn-QSOs are fainter than 19 MACHO $R$ magnitude ($m_R$) while only two of c-QSOs are fainter than 19 $m_R$, which is around a limiting magnitude of MACHO survey (Figure 13). Thus, it is likely that the K-method using variability was not able to detect some of the QSOs due to the large photometric uncertainty and thus weak variability. Figure 14 shows the histogram of the ratio between the average photometric uncertainty and standard deviation (i.e., amplitude), $\sigma/\epsilon$, of the light curves of c-QSOs and cn-QSOs. Small $\sigma/\epsilon$ means that the photometric uncertainty is relatively larger than the amplitude of the light curve, which implies that it is rather hard to detect its variability. As the

**Table 2**
Several Important Parameters of the QSO Candidates

| MACHO ID | R.A. (deg) | Decl. (deg) | $V$ (mag) | hc-QSO[a] |
|---|---|---|---|---|
| 11.8747.1083 | 83.52708 | −70.62689 | 18.98 | 1 |
| 11.8753.346 | 83.66207 | −70.20544 | 18.72 | |
| 11.8984.29 | 83.89623 | −70.89459 | 18.16 | |
| 11.8989.258 | 84.04636 | −70.61672 | 18.67 | |
| 11.8994.1323 | 83.91927 | −70.25463 | 19.27 | 1 |
| 11.9349.1074 | 84.53299 | −70.81329 | 20.26 | 1 |
| 11.9353.1217 | 84.52798 | −70.50399 | 19.52 | |
| 12.10679.528 | 86.51372 | −70.85550 | 18.99 | |
| 13.5834.232 | 79.19451 | −71.16704 | 19.57 | |
| 13.6446.758 | 80.00723 | −70.74329 | 20.32 | |
| 13.6448.3756 | 80.03121 | −70.59326 | 19.04 | |
| 13.6560.555 | 80.25070 | −71.21474 | 19.68 | |

**Note.** [a] 1: high-confidence QSO candidate.

(This table is available in its entirety in a machine-readable form in the online journal. A portion is shown here for guidance regarding its form and content.)

[14] The likelihood ratio between the best-fitting model and a white noise model.
[15] AAT: Anglo-Australian Telescope; AAOmega: the AAT multi-purpose fiber-fed spectrograph (Sharp et al. 2006).
[16] Eighteen of them are on the confirmed MACHO QSO list and seven of them are not on the confirmed MACHO QSO list.
[17] The remaining sources had undetermined classification.
[18] One hundred forty-four newly discovered QSOs and seven previously known QSOs that are not on the confirmed MACHO QSO list.

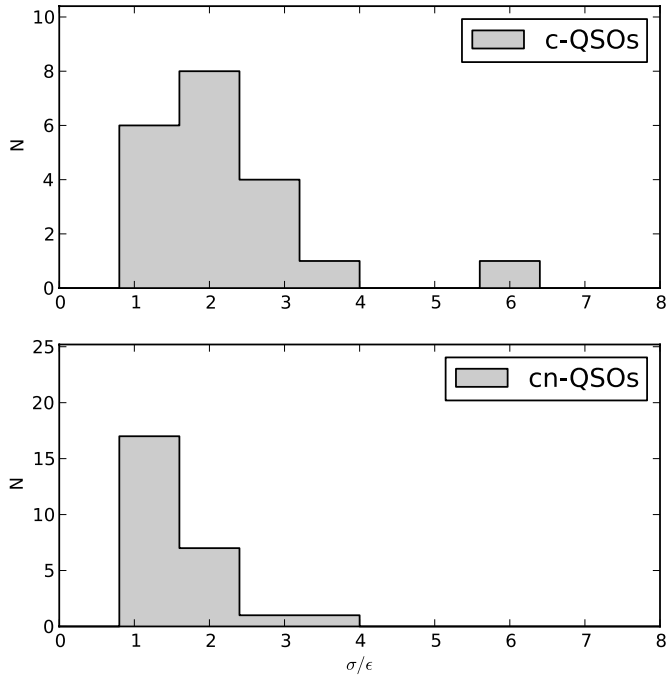


**Figure 14.** Histogram of the ratio between the photometric uncertainty and amplitude, $\sigma/\epsilon$, of c-QSOs (top panel) and cn-QSOs (bottom panel). See the text for details about c-QSOs and cn-QSOs. Small $\sigma/\epsilon$ means that the photometric uncertainty is too large to detect variability. c-QSOs show relatively larger $\sigma/\epsilon$ than cn-QSOs, which means that c-QSOs are more detectable than cn-QSOs using variability.

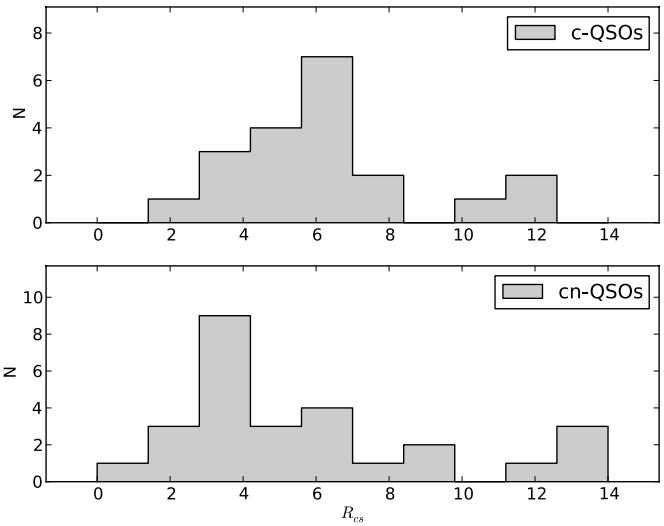


**Figure 15.** Histogram of $R_{cs}$ of c-QSOs (the top label) and cn-QSOs (bottom panel). c-QSOs and cn-QSOs show different distribution. See the text for details.

figure shows, c-QSOs have relatively larger $\sigma/\epsilon$ than cn-QSOs, which means c-QSOs are more detectable than cn-QSOs using their variability. $\sigma$ is one of the time variability features that the K-method used.

In Figure 15, we show an alternative way of seeing variability characteristic of a light curve by borrowing one example of the time series features, $R_{cs}$ (Ellaway 1978), used in the K-method. $R_{cs}$, the range of a cumulative sum, is typically large for the variables showing non-periodic and strong variability, and is small for periodic variables or non-variables. As the figure shows, the histogram of c-QSOs (top panel) has a peak around 6 while the histograms of cn-QSOs shows a peak around 3 (bottom panel).

In addition, we show the MACHO light curves of the 20 c-QSOs and 26 cn-QSOs in Figures 16 and 17. As Figure 16 shows, most of the c-QSOs show strong variability. On the other hand, Figure 17 shows that most of the cn-QSOs fainter than 19 $m_R$ show relatively weaker variability than the variability of c-QSOs. Only cn-QSOs brighter than 19 $m_R$ show strong variability comparable to that of c-QSOs.

According to Figures 14, 15, 16, and 17, it seems that the main reason for the non-detection of QSOs is the relatively weaker variability. Thus, if we ignore some of the QSOs showing weak variability, our efficiency would be higher than 43%. For instance, if we ignore the 11 cn-QSOs fainter than 19 $m_R$, our efficiency increases to 57%.

In the case of the false positives, only 2 out of 122 confirmed non-QSOs are inside the hc-QSO list, which gives 0.3% false positive rate. The two false positives are YSOs. We examined their MACHO light curves and confirmed that they show strong variability. Note that Kozłowski et al. (2012) monitored 12 deg$^2$ fields around the LMC that are mostly inside the 40 deg$^2$ MACHO LMC fields. Given that our QSO candidates are uniformly distributed around the LMC, we would have about one-third of the number of the hc-QSOs (12/40) inside the fields that Kozłowski et al. (2012) monitored. In such a case, the false positive rate is about 1%. However, the true false positive rate would be higher than 1% because Kozłowski et al. (2012) did not monitor all the sources in the fields, which means that some of our QSO candidates are not in their list. Nevertheless, these 122 non-QSOs were selected not only by variability but also by mid-IR colors and X-ray emission. Thus, it seems that our method is successful in excluding any type of false positives, which is crucial for the selection of QSO candidates from massive astronomical databases such as Pan-STARRS (Kaiser 2004) and LSST (Ivezic et al. 2008) due to (1) the enormous amount of data, which thus could yield huge number of false positives, and (2) the high cost of spectroscopic observations for such deep and wide field surveys.

## 6. SUMMARY

In this paper, we presented 663 high-confidence QSO candidates in the LMC fields. We first selected 2566 QSO candidates based on the time variability of MACHO *B*- and *R*-band light curves in the MACHO LMC light curve database using the method of Kim et al. (2011). We then applied multiple diagnostics such as mid-IR color, photometric redshift, and X-ray luminosity to these QSO candidates. Using the diagnostics outputs, we trained a one-class SVM model to discriminate high-confidence QSO candidates. We finally applied the trained model to the original candidates and selected 663 QSO candidates.

To estimate the yield and false positive rate of the final list, we crossmatched them with recently confirmed QSOs and non-QSOs in the LMC field (Kozłowski et al. 2012). As a result, we found that the yield is higher than 43%. It is worth mentioning that the yield of the QSO candidates selected using the "DRW" model (Kelly et al. 2009) is 7% (Kozłowski et al. 2012). In the case of the false positive rate, we found that there are only a few confirmed non-QSOs in our list, which is a less than 1% false positive rate. Thus, this set could be used as a target set potential for spectroscopic survey to maximize the yield. This is important because the spectroscopic observations for relatively faint objects such as the QSO candidates in the dense- and

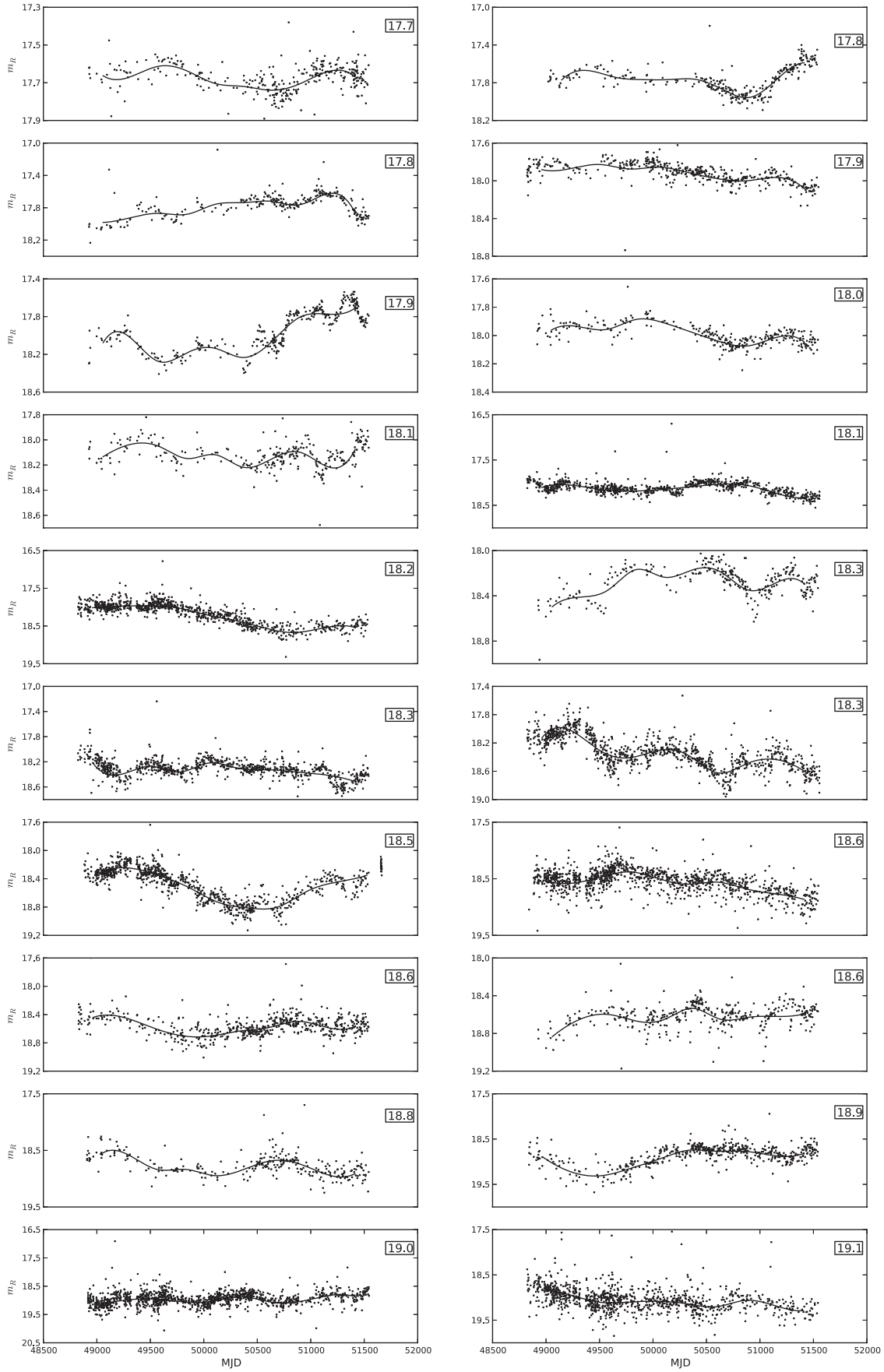


**Figure 16.** MACHO *B*-band light curves of c-QSOs. The *x*-axis is MJD and the *y*-axis is MACHO *R* magnitude ($m_R$). The solid lines are the smoothed spline light curves. The small boxes inside each panel show the average $m_R$. As the figure shows, almost all of the light curves show strong variability regardless of their magnitudes.

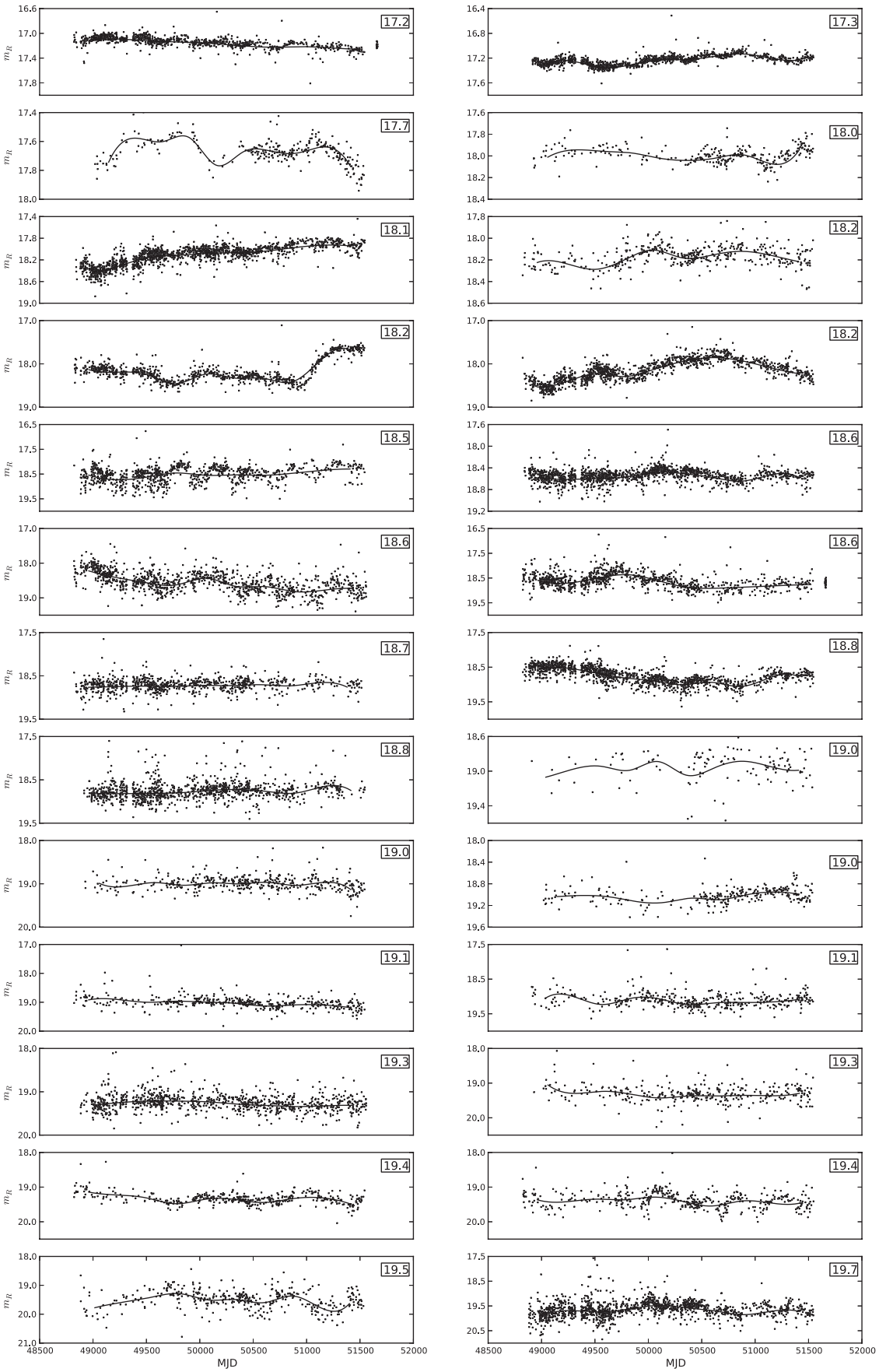


**Figure 17.** MACHO *B*-band light curves of cn-QSOs. When compared to the light curves shown in Figure 16, these light curves show relatively weaker variability. Moreover, there are many more fainter light curves than the light curves in Figure 16.

wide-field area around the LMC is extremely expensive. We are planning to use the confirmed QSOs and confirmed non-QSOs to improve our QSO selection method. This work will be separately published in the near future.

We will apply our method to the MACHO SMC/bulge database and the Pan-STARRS MDF (Medium Deep Field) time series database to further select QSO candidates and thus increase the collection of QSO light curves.

We thank L. Mylonadis for helpful comments. The analysis in this paper has been done using the Odyssey cluster supported by the FAS Research Computing Group at Harvard. This work has been supported by NSF grant IIS-0803409. This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France.

## REFERENCES

Bennett, K. P., & Campbell, C. 2000, SIGKDD Explorations, Vol. 2, 1
Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992, in Proc. Fifth Annual Workshop on Computational Learning Theory, COLT'92 (New York, NY: ACM), 144
Bower, R. G., Benson, A. J., Malbon, R., et al. 2006, MNRAS, 370, 645
Chang, C. C., & Lin, C. J. 2001, LIBSVM: A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm
Cristianini, N., & Shawe-Taylor, J. 2000, An Introduction to Support Vector Machines (Cambridge: Cambridge Univ. Press)
Dobrzycki, A., Eyer, L., Stanek, K. Z., & Macri, L. M. 2005, A&A, 442, 495
Eisenhardt, P. R., Stern, D., Brodwin, M., et al. 2004, ApJS, 154, 48
Ellaway, P. 1978, Electroencephalogr. Clin. Neurophysiol., 45, 302
Elvis, M., Wilkes, B. J., McDowell, J. C., et al. 1994, ApJS, 95, 1
Evans, I. N., Primini, F. A., Glotfelty, K. J., et al. 2010, ApJS, 189, 37
Geha, M., Alcock, C., Allsman, R. A., et al. 2003, AJ, 125, 1
Hawkins, M. R. S. 2002, MNRAS, 329, 76
Heckman, T. M., Kauffmann, G., Brinchmann, J., et al. 2004, ApJ, 613, 109
Hook, I. M., McMahon, R. G., Boyle, B. J., & Irwin, M. J. 1994, MNRAS, 268, 305
Hsu, C.-W., Chang, C.-C., & Lin, C.-J. 2003, A Practical Guide to Support Vector Classification, Technical Report, Department of Computer Science, National Taiwan Univ.
Ivezic, Z., Tyson, J. A., Allsman, R., et al. 2008, arXiv:0805.2366
Kaiser, N. 2004, Proc. SPIE, 5489, 11
Kawaguchi, T., Mineshige, S., Umemura, M., & Turner, E. L. 1998, ApJ, 504, 671
Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, ApJ, 698, 895
Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68
Kollmeier, J. A., Onken, C. A., Kochanek, C. S., et al. 2006, ApJ, 648, 128
Kozłowski, S., & Kochanek, C. S. 2009, ApJ, 701, 508
Kozłowski, S., Kochanek, C. S., Jacyszyn, A. M., et al. 2012, ApJ, 746, 27
Kozłowski, S., Kochanek, C. S., Udalski, A., et al. 2010, ApJ, 708, 927
Lacy, M., Storrie-Lombardi, L. J., Sajina, A., et al. 2004, ApJS, 154, 166
Laurent, O., Mirabel, I. F., Charmandaris, V., et al. 2000, A&A, 359, 887
MacLeod, C. L., Ivezić, Ž., Kochanek, C. S., et al. 2010, ApJ, 721, 1014
Manevitz, L. M., & Yousef, M. 2002, J. Mach. Learn. Res., 2, 139
Meixner, M., Gordon, K. D., Indebetouw, R., et al. 2006, AJ, 132, 2268
Miranda, M., & Macciò, A. V. 2007, MNRAS, 382, 1225
Persic, M., Rephaeli, Y., Braito, V., et al. 2004, A&A, 419, 849
Platt, J. C. 1999, in Advances in Large Margin Classifiers, ed. A. J. Smola et al. (Cambridge, MA: MIT Press), 61
Rees, M. J. 1984, ARA&A, 22, 471
Ross, N. P., Shen, Y., Strauss, M. A., et al. 2009, ApJ, 697, 1634
Rowan-Robinson, M., Babbedge, T., Oliver, S., et al. 2008, MNRAS, 386, 697
Rowan-Robinson, M., Babbedge, T., Surace, J., et al. 2005, AJ, 129, 1183
Schmidtke, P. C., Cowley, A. P., Crane, J. D., et al. 1999, AJ, 117, 927
Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., & Williamson, R. C. 2001, Neural Comput., 13, 1443
Sharp, R., Saunders, W., Smith, G., et al. 2006, Proc. SPIE, 6269, 62690G
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Trichas, M., Georgakakis, A., Rowan-Robinson, M., et al. 2009, MNRAS, 399, 663
Trichas, M., Rowan-Robinson, M., Georgakakis, A., et al. 2010, MNRAS, 405, 2243
Udalski, A., Szymanski, M. K., Soszynski, I., & Poleski, R. 2008, Acta Astron., 58, 69
Watson, M. G., Schröder, A. C., Fyfe, D., et al. 2009, A&A, 493, 339
Zaritsky, D., Harris, J., Thompson, I. B., & Grebel, E. K. 2004, AJ, 128, 1606