Quasar classification and characterization from broadband multi-filter, multi-epoch data sets

Jo Bovy
Institute for Advanced Study / Hubble fellow

Goal

To separate quasars from stars and galaxies using broadband photometric data

To characterize their distances (redshifts)

Quasars—actively accreting supermassive black holes—are among the most luminous objects in the Universe. Large samples of quasars can be used to study topics including inflationary cosmology, the evolution of black hole growth over the course of cosmic history, and the physics of astrophysical black hole accretion. One of the major challenges for the peta-scale surveys of the future is to classify and estimate the distances to quasars without the need for expensive spectroscopic follow-up.

We have broadband photometry available in multiple bands (e.g., *SDSS ugriz*) and in some cases at multiple epochs. This allows us to separate quasars from stars because the colors of quasars are different from those of stars:

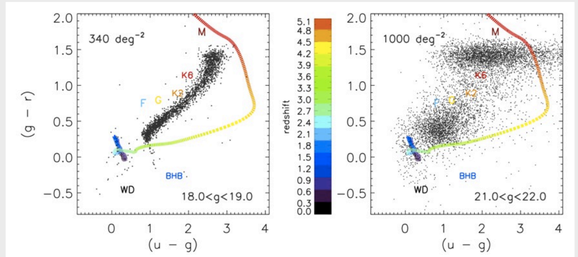


Figure 1: Distribution of quasars (colored locus) and stars (black points) in u-g versus g-r color. The left panel contains relatively bright sources where photometric uncertainties are small and the two loci are separate. The right panel shows fainter sources for which the photometric uncertainties are larger and the two loci overlap. From Ross et al. (2011).

For some stars we also have observations at various epochs. This allows us to use *variability*: quasars are known to vary stochastically on year timescales while stars (and galaxies) do not vary at all, or vary periodically on shorter timescales:

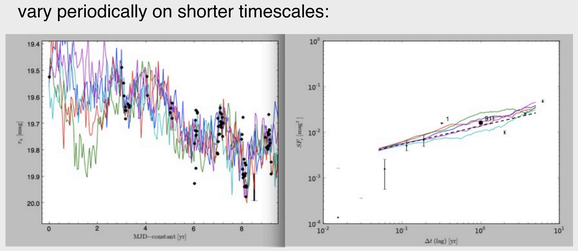


Figure 2: Example lightcurve for a quasar. The left panel shows the r-band lightcurve (black dots), while the right panel shows the structure function defined as:
$V(\boldsymbol{\tau}) = \langle (m(t)-m(t+\boldsymbol{\tau}))^2 \rangle /2 = V(\infty) - C(\boldsymbol{\tau})$, where $C(\boldsymbol{\tau})$ is the correlation matrix
The colored lines are samples from a lightcurve model fit to the observations.

Extreme deconvolution

http://code.google.com/p/extreme-deconvolution/

Summary

Extreme-deconvolution (XD) is a general algorithm to infer a d-dimensional distribution function from a set of heterogeneous, noisy observations or samples. It is fast, flexible, and treats the data's individual uncertainties properly, to get the best description possible of the underlying distribution. It performs well over the full range of density estimation, from small data sets with only tens of samples per dimension, to large data sets with millions of data points.

Likelihood

The observations are assumed to be noisy projections of the true values $\mathbf{v}_i$,

$$\mathbf{w}_i = \mathbf{R}_i \mathbf{v}_i + \text{noise}, \quad (1)$$

where the noise is drawn from a Gaussian with zero mean and known covariance matrix $\mathbf{S}_i$. The case in which there is missing data occurs when the projection matrix $\mathbf{R}_i$ is rank-deficient.

The underlying density is modeled as a mixture of Gaussian components

$$p(\mathbf{v}) = \sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{v}|\mathbf{m}_j, \mathbf{V}_j),$$

then the likelihood is

$$\phi = \sum_i \ln p(\mathbf{w}_i|\theta) = \sum_i \ln \sum_{j=1}^{K} \alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_j, \mathbf{T}_{ij}).$$

We can optimize this likelihood using an *Expectation-Maximization* algorithm

$$\textit{E-step}: \quad q_{ij} \leftarrow \frac{\alpha_j \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_j, \mathbf{T}_{ij})}{\sum_k \alpha_k \mathcal{N}(\mathbf{w}_i|\mathbf{R}_i\mathbf{m}_k, \mathbf{T}_{ik})},$$

$$\mathbf{b}_{ij} \leftarrow \mathbf{m}_j + \mathbf{V}_j \mathbf{R}_i^\top \mathbf{T}_{ij}^{-1}(\mathbf{w}_i - \mathbf{R}_i\mathbf{m}_j),$$

$$\mathbf{B}_{ij} \leftarrow \mathbf{V}_j - \mathbf{V}_j \mathbf{R}_i^\top \mathbf{T}_{ij}^{-1} \mathbf{R}_i \mathbf{V}_j,$$

$$\textit{M-step}: \quad \alpha_j \leftarrow \frac{1}{N}\sum_i q_{ij},$$

$$\mathbf{m}_j \leftarrow \frac{1}{q_j}\sum_i q_{ij}\mathbf{b}_{ij},$$

$$\mathbf{V}_j \leftarrow \frac{1}{q_j}\sum_i q_{ij}[(\mathbf{m}_j - \mathbf{b}_{ij})(\mathbf{m}_j - \mathbf{b}_{ij})^\top + \mathbf{B}_{ij}],$$

Further details can be found in
Extreme deconvolution: inferring complete distribution functions from noisy, heterogeneous and incomplete observations
Jo Bovy, David W. Hogg, & Sam T. Roweis, *Ann. Appl. Stat.* **5**, 2B, 1657 (2011)

XDQSO

We separate quasars from stars by calculating the probability that a given source is a quasar (O \in A below) based on its (single-epoch) fluxes ({a_j} below):

$$P(O \in A|\{a_j\}) = \frac{p(\{a_j\}|O \in A)\, P(O \in A)}{p(\{a_j\})}, \quad (1)$$

where

$$p(\{a_j\}) = p(\{a_j\}|O \in A)\, P(O \in A) + p(\{a_j\}|O \in B)\, P(O \in B), \quad (2)$$

We build models for the probabilities from existing data: confirmed quasars from SDSS-I/II and non-varying point sources (stars). In particular, we deconvolve the *relative-flux* distributions of both quasars and stars using *extreme deconvolution* (see previous section).

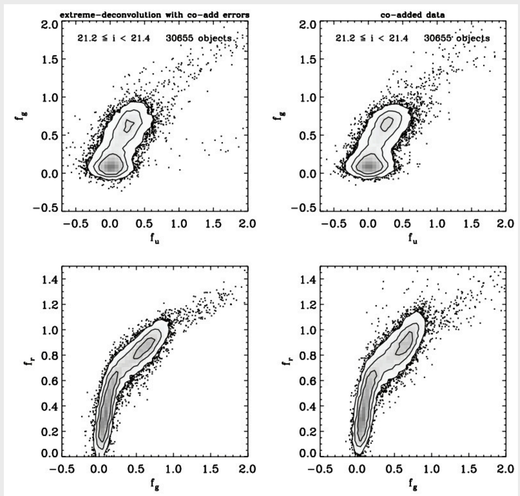


Figure 3: Example deconvolution using XD: distribution of stars in a faint magnitude-bin (right panels) and deconvolved distribution re-convolved with realistic uncertainties (left panels). The underlying distribution is fit using 20 four-dimensional Gaussians.

Using these models for the relative-flux distributions of quasars and stars and using a luminosity-function model for the magnitude distribution of quasars we calculate quasar probabilities for all point-sources in the SDSS DR8 imaging data after convolving the underlying model with the individual uncertainties of the target objects.

Targets are then ranked on probability and the top 20 deg^{-2} are observed as part of *SDSS-III*'s *Baryon Oscillation Spectroscopic Survey*.

After one year of data taking we have discovered approximately 32,000 new quasars, with about 26,000 of those having redshift greater than 4 and an efficiency of about 50 percent. Our approach outperformed other approaches tried (kernel-density estimate, neural net).

XDQSOz

We can use a similar technique to obtain a photometric estimate of the quasar redshift based on its broadband fluxes

By including the redshift in addition to the relative fluxes in our model, we can model the full flux--redshift distribution. This allows us to obtain full redshift probability density functions for all photometric quasars.

These models can be analytically integrated over redshift (as we are using Gaussians as our basic components) to obtain quasar probabilities over arbitrary redshift ranges. The probabilities thus obtained perform as well as the *XDQSO* technique of the previous section.

We also include low signal-to-noise ratio UV data from the *GALEX satellite* and NIR data from *UKIDSS.* These improve both selection as well as redshift estimation. An example is given below

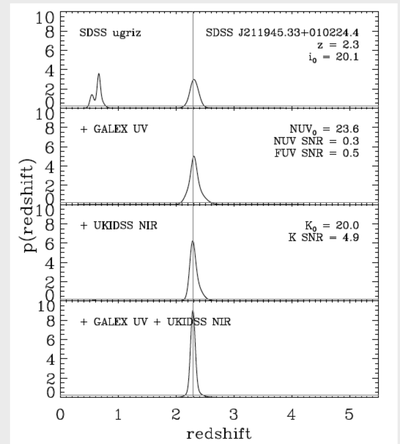
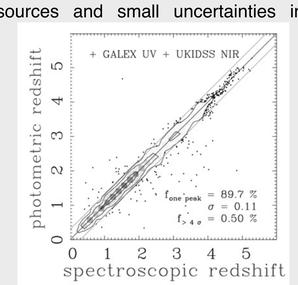


Figure 4: Example photometric redshift PDF and improvement using GALEX UV and UKIDSS NIR data over just using optical SDSS fluxes. The top panel shows the redshift PDF obtained using the XDQSOz flux--redshift models for SDSS optical data. Adding GALEX (second panel) or UKIDSS (third panel) or both (bottom panel) reduces redshift degeneracies. This quasar would be nominally undetected by GALEX and UKIDSS for a 5σ detection limit.

Overall, adding GALEX and UKIDSS we find unambiguous PDFs for most sources and small uncertainties in the photometric redshifts.



References: The work described in this paper is described in
Bovy et al., *ApJ*, **729**, 141 (2011) (*XDQSO*)
Bovy et al., *ApJ*, submitted (2011), arXiv:1105.3975 (*XDQSOz*)
Bovy, Hogg, & Roweis, *Ann. Appl. Stat.* **5**, 2B, 1657 (2011) (*XD*)
Ross, Myers, Sheldon, Yeche, Strauss, Bovy, et al., *ApJS*, in press

Acknowledgements:
Research partially supported NASA (grants NNX08AJ48G and HST-HF-51285.01) and the NSF (grant AST-0908357).

Questions?
email bovy@ias.edu