Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

Automated Software for Slitless Spectroscopy Analysis and Quasar Selection

C. N. Sabbey

Department of Astronomy, Yale University, New Haven, CT 06520-8101

Abstract. The QUasar Equatorial Survey Team (QUEST) is using a 16 CCD drift scan camera on the 1 m Venezuelan Schmidt telescope to collect 30 GB of objective prism data per night. Fully automated software running on a Linux Pentium II reduces and analyses the data faster than it is acquired. Relatively bright spectra in the prism data are automatically detected and used to determine coordinate transformations to the corresponding direct images. The direct image coordinates are used to place spectrum extraction boxes, set the zero point of the wavelength scale, and predict and correct for spectrum overlaps. Multivariate analysis techniques are applied for spectral classification and quasar selection.

1. Introduction

The QUEST collaboration (Snyder 1998) is obtaining broad bandpass (4000 – 9500Å), intermediate dispersion (25 Å/arcsec at Hα) slitless spectroscopy covering 300 deg^2 per night to $m_B \approx 19$. Processing the $\sim 10^6$ spectra contained in the nightly 30 GB of image data, however, places unique demands on the analysis software. The software must be fast and fully automated, reliably extract, deblend, and wavelength calibrate spectra that are often not detectable until later coaddition, and take advantage of the broad bandpass (continuum information) to improve quasar selection.

2. Coordinate transformation and spectrum extraction

By using a direct image catalog of sky coordinates and magnitudes, the locations of the spectra and their contamination by neighboring objects can be predicted. We currently use a catalog generated from the POSS/STScI Digitized Sky Survey (DSS), with the help of the Sextractor (Bertin & Arnouts 1996) and WCSTools (Mink 1997) software packages.

Relatively bright spectra in the prism data are automatically detected and their coordinates are input to a triangle matching program (Valdes et al. 1995) to determine the linear coordinate transformation to the direct catalog. The spectrum position in the dispersion direction is defined by the point of half maximum brightness at the red end sensitivity cutoff. Although this definition will depend on spectral type (by ± 2 pixels), we use it to establish the mean offset and not to set the zero point on an individual basis.

The relatively bright spectra are automatically detected by convolving the image with a matched filter (a 2-D template of the red end cutoff), and then searching for vertical strings of relative maxima that monotonically increase in brightness. Strings of some minimum length (e.g., ten pixels, possibly diagonally connected) reliably locate the distinctive red end cutoff and hence the entire spectrum (which is $\approx$ 250 pixels long). Object detection algorithms used in photometry (i.e., connecting pixels above a brightness threshold) are less useful due to the extreme overlap problem and the segmentation of an object into multiple detections depending on the spectral energy distribution.

The software applies the coordinate transformation to the direct catalog, predicting the spectrum locations to within $\approx 0\farcs3$ (except near photographic plate edges in the DSS data). This accuracy is required to reliably extract faint spectra and set the wavelength zero point to within $\approx$ 0.2%. The accuracy of the transformation is monitored throughout the drift scan by using the ≈ 50 brightest spectra per $0\fdg6 \times 0\fdg6$ image frame as "guide stars" (least squares fits to the position residuals refine the transformation). The extraction of 1-D spectra is currently done using a box of width $1.2 \times$ FWHM and uniform weighting.

3. Background determination

The background level is determined by a technique commonly used for crowded fields — image histogram mode estimation with iterative k-sigma clipping of the histogram wings (Da Costa 1992). A fast and straightforward method for implementing this uses the cumulative histogram, `c`, and the weighted cumulative histogram `w` (i.e., the contents of each histogram bin are first multiplied by the bin number). At each iteration, the mean, 25th percentile (`q25`), median (`q50`), mode, and sigma (`sig`) are obtained for the clipped histogram (i.e., between bins `b` and `e`) as follows:

```
nincr = c[e] - c[b-1]
mean  = (w[e] - w[b-1]) / nincr
q25   = ibsearch(c[b-1] + nincr / 4)
q50   = ibsearch(c[b-1] + nincr / 2)
mode  = 3 * q50 - 2 * mean
sig   = 1.482 * (q50 - q25)
```

where `ibsearch` indicates a binary search of the cumulative histogram with an interpolated output index. Convergence (when the mode changes by less than some threshold) occurs rapidly by using `q25` and `q50` to form a robust approximation of the histogram σ. This method is $\sim$ 10 times faster than techniques that require fully sorting the data. First masking out the pixels within the spectrum extraction boxes did not improve the background estimation.

4. Overlap prediction and correction

A spectrum is considered to be overlapped if a significant number of the wavelength bins have more counts due to neighboring spectra in the image than the spectrum itself. Whether a given spectrum is overlapped depends on the relative magnitudes and positions of the neighboring objects, their spectral types and the spatial profile. In the QUEST objective prism data, almost 30% of the 19th magnitude spectra have $\geq 25\%$ of their bins dominated by neighbors. However, the software predicts the amount of contamination as a function of wavelength to attempt deblending and allow the unrestricted use of clean wavelength bins.

The software builds a contamination map using the positions and magnitudes of objects in the input sky catalog, and empirical 2-D template spectra produced from bright, uncrowded spectra. Currently a single, representative late type spectrum is used although an improvement would be to create templates as a function of color. The template spectra are added into the map with an intensity scaling of $10^{(m/(-2.5))}$. For each spectrum extracted from the data, the ratio of the contamination map to the 2-D template for that spectrum gives the (multiplicative) contamination correction at each pixel.

This method is fast, easy to implement, and handles arbitrarily complicated overlaps of many objects without excessive approximation. In addition, the overlap calculation can be done independently of the spectrum extraction pipeline. Thus, the overlap calculation does not need to be repeated for each drift scan to be coadded, and the calculation can be improved at any time without repeating the spectrum extraction. Because the input sky catalog contains ID numbers for all objects to be extracted, it is straightforward to match each spectrum to its contamination array (and coadd spectra from different CCDs and nights).

5. Quasar selection

With the broad bandpass of the spectra, spectral classification and quasar selection can be improved by using the continuum information in addition to spectral features. A reasonable approach is to find the smallest RMS deviation between each spectrum and a series of template spectra (or similar technique based on cross-correlation). However, two significantly different spectra can have the same RMS deviation from a given template (e.g., one spectrum has an emission feature while the other has a slightly different continuum slope). In other words, the best matching template is selected using a distance criterion, independent of direction, in flux–wavelength space.

Figure 1 shows an alternative using Principal Component Analysis (PCA) to provide a fast, automated measure of the deviation of a given spectrum from the common energy distributions. This allows the selection of quasars (and other unusual objects) to fainter magnitudes than line-only searches without making assumptions about the variance of quasar spectra. In practice, the representative set of common spectra is defined from the data itself (by clustering analysis on a sample of relatively bright, unblended spectra).

References

Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393

Cristiani, S. & Vio, R. 1990, A&AS, 227, 385

Da Costa, G. 1992, ASP Conf. Ser., Vol. 23, Astronomical CCD Observing and Reduction Techniques, ed. S. Howell (San Francisco: ASP), 90

Gunn, J. E. & Stryker, L. L. 1983, ApJS, 52, 121

Mink, D. 1997, in ASP Conf. Ser., Vol. 125, Astronomical Data Analysis Software and Systems VI, ed. G. Hunt & H. E. Payne (San Francisco: ASP), 249

Snyder, J. A. 1998, in SPIE Proc., Vol. 3355, Optical Astronomical Instrumentation, ed. S. D'Odorico (Bellingham: SPIE), 635

Valdes, F., Campusano, L., Velasquez, J., & Stetson, P. 1995, PASP, 107, 1119

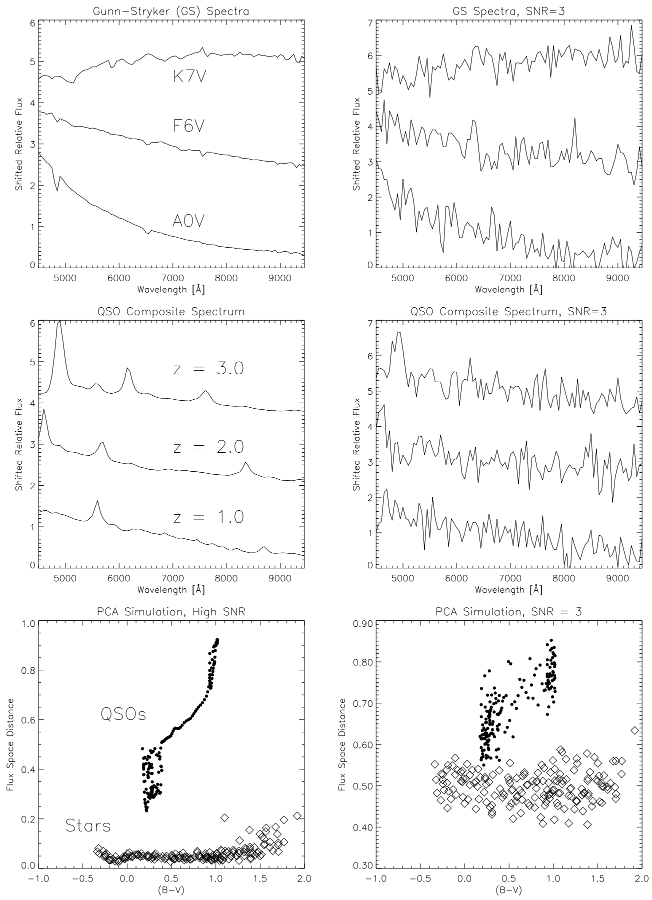

Figure 1. A simulation is shown using PCA to select uncommon (non-stellar) energy distributions. The Gunn & Stryker (1983) stellar spectra and a composite QSO spectrum (Cristiani & Vio 1990) for $0.5 < z < 4.5$ are rebinned to 50 Å bins in the 4500 to 9500 Å bandpass. PCA is applied to the GS spectra to establish a new basis in which the first 5 dimensions account for $> 99\%$ of the stellar spectral variance. A measure of the deviation of a given spectrum from stellar is obtained by projecting it into the new basis and calculating the flux space distance to the volume formed by the N (e.g., 5) low order dimensions.