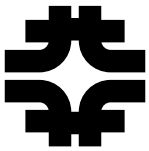


Fermi National Accelerator Laboratory

FERMILAB-Pub-97/014

Three Dimensional Parameterization of the Stellar Locus with Application to QSO Color Selection

Heidi J. Newberg and Brian Yanny

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

January 1997

Submitted to *Astrophysical Journal*

Operated by Universities Research Association Inc. under Contract No. DE-AC02-76CH03000 with the United States Department of Energy

Disclaimer

Distribution

FERMILAB-Pub-97/014

Three Dimensional Parameterization of the Stellar Locus with Application to QSO Color Selection

Heidi J. Newberg and Brian Yanny
Fermi National Accelerator Laboratory, Box 500, Batavia, IL 60510
Electronic mail: heidi@fnal.gov, yanny@fnal.gov

ABSTRACT

A straightforward method for parameterizing and visualizing a locus of points in n-space is presented. The algorithm applies directly to the problem of distinguishing QSOs from Galactic stars in multicolor space. When applied to an existing four-filter data set (photometric errors of $\sigma \sim 0.1$ mag and $B_J \lesssim 21.5$) it recovers 95% of the known QSOs, with 76% of the candidate sources yielding QSOs. The algorithm scales well to large data sets, and allows astrophysical information about the location of QSOs in color space to be easily added to improve efficiency. Three-dimensional visualization of the high accuracy photometry ($\sigma < 0.01$ mag) of the *Catalogue of WBVR Magnitudes of Northern Sky Bright Stars* (V<7.2) yields the first look at the intrinsic width in the Galactic star color distribution. In the type G to early type K star region, the cross section of the stellar locus has a FWHM of 0.07 mag. This width is not due to reddening or measurement errors, but reflects an intrinsic property of the stellar locus. The algorithm can be applied to parameterize any one-dimensional set of data that is distributed in n-dimensional space.

Subject headings: quasars, stars:general, stars:statistics

1. Introduction

Quasi-stellar objects (QSOs) are believed to be the observable manifestations of black holes at the centers of distant galaxies (Hazard 1979). Large samples of QSOs are important for studying the intrinsic properties of the objects as well as their space distribution and evolution. As the most distant observable collapsed objects in our universe, QSOs constrain cosmological models and contribute to our knowledge of the matter distribution on large scales and at early times. QSOs are also used as "flashlights" which illuminate the non-luminous matter in their line of sight.

Unlike galaxies, QSOs cannot be selected from single images of the sky, since they are indistinguishable from Galactic stars. Techniques for identifying QSOs include looking at radio

sources, x-ray sources, objective prism spectra, objects with non-stellar colors, variable objects, or objects with lack of proper motion. For most of these techniques, follow-up spectroscopy is required for positive identification and determination of redshift. Each technique has its own advantages and selection biases; several search modes must be used in combination to produce a full picture of the variety of QSOs in the universe. In this paper, we discuss a new algorithm for the selection of QSOs with non-stellar colors.

Color selection of QSO candidate objects with an ultraviolet excess (UVx) has been used successfully for almost as long as QSOs have been identified (Braccesi, Lynds, and Sandage 1968, Janes and Lynds 1969, Green, Schmidt, and Liebert 1986). Until recently, however, few researchers have selected their QSO candidates using more than two colors at a time, even if more than two colors were available at the time the selection was made (Braccesi, Formiggini, and Gandolfi 1970, Zhan, Koo and Kron 1989, Koo and Kron 1982, Hall et al. 1996). Those who have used more than two colors at the same time have based their selection on n^{th} nearest neighbor distances (Warren et al. 1991a), or distance from the stellar locus (Gaidos, Magnier and Schechter 1993) or both (Kron et al. 1991).

The nearest neighbor distance criterion (Warren et al. 1991a) chooses those objects which are in the lowest density regions of magnitude space. In this technique, the magnitude in multiple filters of every source in the catalog is compared with every other source to measure two metric distances for each source. The first distance is the simple Euclidean distance to the n^{th} (where n is a small integer like 10) nearest source in magnitude space. The second is this distance weighted by the photometric error, which produces the number of standard deviations that separate them. Candidate QSOs are chosen by setting a threshold in a linear combination of these two density measures.

This nearest neighbor distance criterion, which measures the source density in magnitude space, is more efficient at selecting QSOs than selection based on a set of two dimensional plots (Hall et al. 1996). However, there are scientific drawbacks to the method. For instance, the color density of UVx quasars can be large enough in some regions of color space that those quasars may be rejected as clumps of stars (Kron et al. 1991). Also, the relative densities of stars and quasars change as a function of magnitude and field location (Kron et al. 1991), making the selection criteria difficult to understand and causing spurious candidates at brighter magnitudes (Warren et al. 1991a, Porter et al. 1991). In addition, the procedure can be difficult to implement for large samples with millions of objects. A traditional $O(N^2)$ nearest neighbor algorithm, in which every source is compared with every other source, is too slow to use with large surveys such as the Sloan Digital Sky Survey ($N \sim 10^8$ stellar objects) (Gunn 1995). One may be able to reorganize the data to reduce the order of the algorithm, but in general, finding the k-nearest neighbors of N points is a time-consuming process for large N (Fukunaga and Narendra 1975).

An extension of the nearest neighbor technique, which addresses many of the previously mentioned problems, is to divide n-dimensional magnitude space into boxes with size approximately

equal to the errors in a given data set. Then, we can measure the stellar density and the QSO density in each of these boxes and thus compute the probability that a source in that box in color space is a QSO. We would select as candidates those sources for which the probability of being a QSO is greater than some threshold. Of course, it would be necessary to compute these probabilities as a function of Galactic latitude and longitude. Although this statistical approach is theoretically feasible and would result in optimal QSO selection, computing the densities would require too many sources of known type for this to be practical.

Instead of the purely statistical approach, we have chosen as inspiration the *distance from the stellar locus* criterion and the algorithm of Gaidos, Magnier and Schechter 1993. In this algorithm, all point-like sources are plotted in multicolor space, and then a set of locus points is fit along the center of the stellar locus. The QSO candidates are those that are further than some metric distance away from the locus fit. Gaidos, Magnier and Schechter 1993 use both the Euclidean distance in multicolor space and an error weighted distance.

Like Gaidos, Magnier and Schechter 1993, we generate a small, finite set of locus points that describes a line fit through the center of the stellar locus. We go further in parameterizing the locus, however, by measuring the perpendicular widths in the thick and thin directions for each locus point along the stellar locus. Therefore, we do not select QSOs candidates as sources that are more than a fixed *distance* from the stellar locus, but rather as those objects that are outside a manifold of elliptical cross section. This allows us to detect and quantify the intrinsic width of the locus of stellar sources. Our technique begins with the well-defined problem of parameterizing the stellar locus, so that objects within it can be excluded from consideration as QSO candidates. The algorithm can then be incrementally improved using astronomical, rather than statistical, considerations.

2. Parameterization of the stellar locus

2.1. Goals of the algorithm

Our goal is to start with a set of colors of stellar sources, and from that produce a set of locus points that run down the center of the locus of stars. The algorithm to fit the stellar locus is currently implemented for three dimensional multicolor space, but the technique can be extended to arbitrary dimension, as long as the locus remains approximately 1-D (a line) in n-space. We will use *stellar locus* to refer to the data points in multicolor space, and *locus points* to refer to the model fit. At each locus point, we will compute a unit vector $\hat{k_p}$, in the direction along the stellar locus. Also at each locus point, we will attempt to describe the distribution of stars in the plane perpendicular to the unit vector $\hat{k_p}$. We do this by fitting an ellipse to the cross section of the stellar locus. The ellipse is oriented so that the major axis is in the direction that the stellar locus is the widest, and the minor axis is in the direction that the stellar locus is the narrowest. The sizes of the major and minor axes are computed from the width of the distributions of stars

in their respective directions.

The parameters produced from this algorithm describe not only where the stellar locus lies in color space, but also how its width changes as a function of distance along the locus. A powerful result of this parameterization is that we can set up local coordinate systems along the locus that allow us to visualize the data in ways that were not previously possible. We can assign a distance along the locus, distance away from the center of the locus in what is locally the thickest direction, and distance away from the center of the locus in what is locally the thinnest direction. Using these coordinates, we can now plot slices through the locus at any point, or densities of sources along the locus. These plots are useful for understanding the data, and the parameterization is useful for defining outliers from the parameterized distribution.

Section 2.2 gives a more rigorous overview of how the algorithm arrives at a solution. Sections 2.3-2.7 give the details and mathematical foundations on which the algorithm is built, with an outline of how the algorithm can be extended to higher dimension. Section 2.8 describes outlier rejection, an optional last step that in many cases will improve the accuracy of the final solution. Section 2.9 discusses limitations and possible extensions of the algorithm. Section 2.10 discusses how the algorithm can be extended to accommodate individual errors for each of the sources in the input catalog.

2.2. Overview of the algorithm

Figure 1 shows a graphical sketch of how the algorithm progresses towards the solution. The data used for this demonstration is from Bershady et al. 1997. The algorithm starts out by placing two locus points at the coordinates it has been given as the approximate endpoints of the distribution. These locus points, unlike the endpoints, are not at a fixed position, but will move as the algorithm progresses. As we iterate through the algorithm, new locus points will be added and the positions of all of the locus points will be adjusted.

During each iteration, each locus point is moved to the centroid of its associated stars. For illustrative purposes, suppose that the associated stars are those that are closer to that locus point than they are to any other locus point. Then, $\hat{k_p}$ and an ellipse fit are calculated for each locus point. With each iteration, new locus points are added in between the existing locus points, but only in places along the locus where the distance between the locus points is larger than a factor times the local width of the stellar locus. We iterate until the algorithm converges.

2.3. Input parameters

The concept of a best curve fit through the center of a locus of stars is not rigorously defined. In other words, if there are two independent fits to the curve, there is no unbiased procedure

to determine which fit is more correct. For example, imagine we have a cross section through the stellar locus (itself not a well defined entity). Should we place a locus point in the densest region? Should we use the average position of the group of stars? Should we reject outliers, such as reddened stars, variable stars, or binaries? If the locus broadens out substantially (as it does at the red end), how much freedom do we give the locus fit to wind through structure in the locus of stars? How are the endpoints of the stellar locus determined?

Given the difficulty in defining the best-fit locus points, we include several input parameters which can be adjusted to give an acceptable fit to the stellar locus for a particular set of data points. These parameters are: the two approximate endpoints of the stellar locus, $\vec{r}_{start}$, $\vec{r}_{end}$; the maximum distance from the stellar locus (d_x, d_y, d_z) that a star will still be considered as part of the locus; the number of iterations of the algorithm N_{iter}; and the factor, $N_{\sigma_{spacing}}$, which determines how closely spaced the locus points can be.

The endpoints of the stellar locus can be simply determined from 2D plots of the data. On the first iteration of the algorithm, two locus points will be located at the two endpoints, so they should not be chosen to be farther from the mass of stars than the maximum distance we will look for associated stars. Note that we use the term *endpoints* to refer to the input parameters. The line fit through the center of the locus of stars is not constrained to go through these endpoints. Each end of the stellar locus fit will ultimately be defined by a plane that goes through the endpoint and is normal to the line which connects the first two (or last two) locus points.

The maximum distances that associated stars can be from a given locus point must be large enough so that when points are added half way between existing locus points, the new point will have enough associated stars that it can migrate to the locus. The associated stars of locus point p lie between two parallel planes which are perpendicular to $\hat{k_p}$ and are half way between locus point p and the adjacent locus points. In some cases, the locus curves enough that stars in a distant part of the color space will be associated with a given locus point. The maximum distances must be chosen to be small enough that this does not happen.

The accuracy with which a locus point fits the center of the distribution of stars depends on several things, including the local density of stars and the intrinsic width of the stellar locus. If we allow the locus points to move arbitrarily close to each other, they can become closer than the errors in their positions, causing the algorithm to become unstable. To prevent this, we refrain from adding new locus points between existing locus points that are closer than $N_{\sigma_{spacing}}$ times the major axis of the ellipse fit at that point in the locus. We typically use $N_{\sigma_{spacing}} = 3$. We continue iterating until the number of locus points is limited by the width of the locus itself, and then iterate several more times until the locus has stabilized.

2.4. Locus cross section

When the iteration starts, there are only two locus points. The first task is to generate ellipse fits to the cross section of the locus at each of those points, so we can decide whether to add another point between them.

First, we must compute the set of unit vectors, $\hat{k_p}$, that point along our set of locus points. The direction of $\hat{k_p}$ is from the previous locus point to the next locus point:

$$\hat{k}_p \equiv (\vec{r}_{p+1} - \vec{r}_{p-1})/\|\vec{r}_{p+1} - \vec{r}_{p-1}\|,$$

where the $\vec{r}_p$ are the positions of the locus points in three dimensional color space. In the case that the locus point is the first or the last, the unit vector is determined from itself and its adjacent locus point.

Next, we look at each star and determine with which locus point(s) it is associated. A star at position $\vec{r}$ is associated with locus point p if it is between the plane $(\vec{r} - (\vec{r}_p + \vec{r_{p+1}})/2) \cdot \hat{k}_p = 0$ and the plane $(\vec{r} - (\vec{r}_{p-1} + \vec{r_p})/2) \cdot \hat{k}_p = 0$, and $\|(\vec{r} - \vec{r}_p) \cdot \hat{x}_i\| < d_{x_i}$ for all $x_i \in \{x, y, z\}$. In other words, a star is associated with a given locus point if it is between two planes with normal vectors along the locus that intersect the midpoints between the given locus point and the two adjacent locus points, and if it is also closer to the locus point than the maximum distance parameters in each of the coordinates. If the given locus point is the first (or last) locus point, then the beginning (or end) endpoint is used instead of the midpoint between the current and previous (next) locus point.

The reason that we do not define a locus point's associated stars to be those stars that are closer to the locus point than they are to any other is that this definition does not result in a stable algorithm. Imagine we have a straight and cylindrically symmetric stellar locus. If one of the locus points is slightly perturbed, its associated stars will lie inside a wedge shape in color-color-color space. The thick end of the wedge (where the majority of the stars are) is on the same side of the center of the locus as the perturbation, causing the locus point to move even farther from the center of the locus. Although our definition of the associated stars will tend to move the locus points towards the inside of the curves of the stellar locus, it is preferred over an unstable solution.

We define a local coordinate system $< \hat{i_p}, \hat{j_p}, \hat{k_p} >$ around each locus point. Here, $\hat{k_p}$ is the previously computed unit vector along the locus, and:

$$\hat{k_p} \equiv k_x\hat{x} + k_y\hat{y} + k_z\hat{z}$$

$$\hat{j_p} \equiv (\hat{k_p} \times \hat{z})/|\hat{k_p} \times \hat{z}| = (k_y\hat{x} - k_x\hat{y})/\sqrt{k_x^2 + k_y^2},$$

$$\hat{i_p} \equiv \hat{j_p} \times \hat{k_p} = (-k_xk_z\hat{x} - k_yk_z\hat{y} + (k_x^2 + k_y^2)\hat{z})/\sqrt{k_x^2 + k_y^2}.$$

Note that $\hat{j_p}$ is always perpendicular to the third ($\hat{z}$) color coordinate axis. For each source s associated with a given locus point, we compute the vector $\vec{d_s} = i_s\hat{i} + j_s\hat{j} + k_s\hat{k}$ which goes from that locus point to the source in color-color-color space.

The next step is to calculate the ellipse (defined by a_p, b_p, and θ_p) associated with each locus point. Since we have divided the locus up into segments along the $\hat{k_p}$ direction, we need only consider the distribution in the $\hat{i_p}, \hat{j_p}$ plane. We define two new coordinates:

$$\hat{l_p} \equiv \cos\theta_p \hat{i_p} + \sin\theta_p \hat{j_p},$$

$$\hat{m_p} \equiv -\sin\theta_p \hat{i_p} + \cos\theta_p \hat{j_p},$$

where $-\pi/2 < \theta_p \leq \pi/2$. θ_p is defined to be the angle between the major axis of the ellipse and the unit vector $\hat{i}$. We will force $\hat{l_p}$ to point along the major axis of the best fit ellipse by minimizing $\sum_s m_s^2$ where $m_s = \vec{d_s} \cdot \hat{m_p} = -\sin\theta_p i_s + \cos\theta_p j_s$. The solution of this minimization is:

$$\tan\theta_p = \begin{cases} 0 & M_{ij} = 0, M_{ii} > M_{jj} \\ \infty & M_{ij} = 0, M_{jj} > M_{ii} \\ \frac{-(M_{ii}-M_{jj})+\sqrt{(M_{ii}-M_{jj})^2+4M_{ij}M_{ij}}}{2M_{ij}} & M_{ij} \neq 0 \end{cases}$$

where the second moments of the distribution are given by:

$$M_{ii} \equiv \sum_{s=1}^{N_p} \frac{i_s^2}{N_p},\ M_{ij} \equiv \sum_{s=1}^{N_p} \frac{i_s j_s}{N_p},\ \text{and}\ M_{jj} \equiv \sum_{s=1}^{N_p} \frac{j_s^2}{N_p}.$$

Here, N_p is the number of associated stars for locus point p. The major and minor axes of the ellipse shall be defined as:

$$a_p \equiv \sigma_l = \sqrt{\frac{\sum_s l_s^2}{N_p}} = \sqrt{\frac{M_{ii} + M_{jj} + \sqrt{(M_{ii} - M_{jj})^2 + 4M_{ij}M_{ij}}}{2}}$$

$$b_p \equiv \sigma_m = \sqrt{\frac{\sum_s m_s^2}{N_p}} = \sqrt{\frac{M_{ii} + M_{jj} - \sqrt{(M_{ii} - M_{jj})^2 + 4M_{ij}M_{ij}}}{2}}$$

A method to determine the axes of the ellipse which more easily generalizes to higher dimension would be to define the $(\hat{l}, \hat{m})$ coordinate system as the coordinate system which diagonalized the matrix of second moments. In this case, the eigenvectors of the second moment matrix define the directions of the $\hat{l}, \hat{m}$ vectors, and the sigmas are the square roots of the eigenvalues. This allows us to extend the algorithm to higher dimension by using the eigenvalues and eigenvectors of an $(n-1) \times (n-1)$ matrix to determine the axes and widths of the $(n-1)$-dimensional ellipsoids associated with each locus point.

2.5. New locus points

For each set of two adjacent locus points in the current set, we decide whether to add a new locus point bisecting the line segment between them. A new point is added if the Euclidean

distance between the two locus points is larger than $N_{\sigma_{spacing}}$ times the average of the major axis fits of those locus points. This feature of increasing the spacing of locus points where the locus is broad is a unique feature of this algorithm, and is one reason we achieved a stable fit for the locus points.

2.6. Centroid of stellar locus

The next step is to move each locus point to the centroid of its associated stars. First, the associated stars for each locus point are calculated in exactly the same way as they were in §2.4. Then, the position of the locus point is moved to the average position of all of these associated stars. From this point, we loop back to adding new locus points until we have fulfilled all of the iterations of the algorithm.

2.7. When the Algorithm Finishes

After the last iteration, we recompute the quantities a_p, b_p, θ_p, and $\hat{k_p}$ for each of the final locus points. We also compute the coordinate along the locus for each of the locus points using:

$$k_p \equiv \begin{cases} 0 & p = 1 \\ \sum_{i=2}^{p} |\vec{r}_i - \vec{r}_{i-1}| & p > 1 \end{cases}$$

2.8. Outlier rejection

In some cases, there are a significant number of sources that are not part of the central distribution of stars. These could include variable stars, binaries, quasars, or reddened stars. Since we do not want these to pull the locus points away from the central distribution of stars, we implemented a separate routine that will recalculate the locus points rejecting outliers of the distribution. This routine iterates a specified number of times, each time rejecting outliers that are $N_{\sigma_{reject}}$ sigma away from the center of the distribution.

First, a new centroid is calculated from the associated stars of each locus point. The associated stars are defined as above, with the additional constraint that $l_s < N_{\sigma_{reject}} a_p$, and $m_s < N_{\sigma_{reject}} b_p$. This criterion includes a rectangular area in the (l, m) plane. The unit vectors along the locus, $\hat{k_p}$, are recalculated using the new locus point centers.

Since the locus points have now moved, we must re-determine the associated sources for each locus point, defined with outlier rejection. Using the new associated stars, the values of a_p, b_p, and θ_p are calculated from the second moments as before. Since these sigmas were calculated without including the "outliers", they will be underestimates of the true sigmas of the distribution. We correct for the underestimate assuming that the underlying distribution is Gaussian. One can

calculate that:

$$\sigma^2_{clipped} \equiv \frac{\int_{-N\sigma}^{N\sigma} x^2 A e^{-\frac{x^2}{2\sigma^2}} dx}{\int_{-N\sigma}^{N\sigma} A e^{-\frac{x^2}{2\sigma^2}} dx} = \sigma^2 \left\{ 1 - \frac{N e^{-N^2/2}}{\sqrt{2\pi}\{F(N) - \frac{1}{2}\}} \right\}.$$

where

$$F(N) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{N} e^{-\frac{x^2}{2}} dx.$$

Therefore, we estimate the true sigma by dividing the clipped sigma by the factor:

$$\text{clip factor} = 1 - \frac{N e^{-N^2/2}}{\sqrt{2\pi}\{F(N) - \frac{1}{2}\}}.$$

We then go back to calculating new centers for the locus points using the new locus parameters, and complete the requested number of outlier rejection iterations.

2.9. Enhancements and Limitations of the Algorithm

To summarize, we have created a set of ordered locus points together with the direction of the locus and an ellipse fit to the cross section of the locus at each of these points. By parameterizing the stellar locus in this way, we have essentially performed a non-linear principal components analysis. Each star can be associated with its closest locus point, and assigned a value of l_s, m_s, and k_s using the local $(\hat{l}, \hat{m}, \hat{k})$ coordinate system as described in section 2.1. In sections 3 and 4 we will demonstrate the advantages of this parameterization by using these coordinates to plot cross sections of the locus (m_s vs. l_s) between specified values of k (along the locus), and histograms showing the stellar density along the locus. We can use the stellar density along the locus to set the blue endpoint of the locus (the locus generally cuts off more sharply on the blue end than on the red end), or to identify the 3D location of some other feature. Once the position of the locus can be associated with a set of points, one can easily determine an offset between the stellar loci of different data sets. The parameterization also allows us to select outliers (such as color selected QSOs) from the distribution in several ways. In section 4, we use the most straightforward approach: we choose as candidates any sources that are more than $N_{\sigma_{select}}$ sigma events, where sigma is given by the major and minor axes of the ellipses associated with each locus point, and $N_{\sigma_{select}}$ is a small number like 4. Alternatively, one can redistribute the locus points so that there are about N_s sources associated with each locus point, and then rescale the major and minor axes of the ellipse fits at each of the new locus points so that $x\%$ of the associated stars will be contained within the fit ellipse (here, N_s and $x\%$ are input parameters). One could imagine selection schemes which make use of the density, as measured from a histogram along the locus, to change $N_{\sigma_{select}}$ as a function of position along the locus. This would allow the use of the local density of sources as a cutoff, going deeper into the stellar locus where there are fewer stars. Or, one could lower $N_{\sigma_{select}}(k)$ along the locus to favor regions that are known to contain more QSOs (digging deeper at the blue end).

We also have the option to include additional selection criteria. For example, one might find that in addition to the fit stellar locus, which includes the majority of Galactic stars, there is a locus of metal-poor blue stars or compact blue galaxies that can be easily excluded. Or one could exclude the red end of color space if no quasar has been found, or is expected to be found, in that region.

The algorithm as currently implemented has some limitations. For example, the current implementation of the algorithm will not work for a set of points that are arranged in the shape of a horseshoe. In order to keep the points on one extreme of the horseshoe from affecting the centroid of the other extreme of the horseshoe, the input parameters that control the maximum distance from the "stellar" locus that a "star" would be considered part of the "stellar" locus (d_x, d_y, d_z) must be set smaller than the distance between the two endpoints of the horseshoe distribution. However, the first time the algorithm iterates, a locus point will be put exactly between the two endpoints of the distribution in such a position that it will have no associated "stars" at all. Depending on the exact situation, the algorithm may not converge properly. A simple solution is to specify more than two initial locus points, and not require that these points be at the same position as the endpoints. One can then space the points around the horseshoe closely enough to allow the algorithm to converge properly.

Although it is straightforward to generalize the algorithm to work in n-space, we have coded only the $n = 3$ case. Coding the more general algorithm requires the ability to find eigenvalues and eigenvectors of $(n-1) \times (n-1)$ matrices, as mentioned in §2.4. Two dimensional data can be fitted with the current implementation of the algorithm by using a fixed constant for the third coordinate of all of the input data points.

It is not guaranteed in the current implementation of the parameterization that two points that are close in color-color-color space will be close in (l, m, k) space. It is also possible for two points in color-color-color space to be mapped into the same coordinate in (l, m, k) space. This is because the (l, m, k) coordinates are assigned based on the parameters associated with the closest locus point, which makes the mapping discontinuous. One could imagine using the locus points and their associated parameters to compute a smooth, continuous transformation from color-color-color space to (l, m, k) space, but this work has not been done.

2.10. Handling errors on individual data points

The most obvious unimplemented feature of our algorithm is the ability to recognize different errors for different data points. Ideally, one wants to give less weight to stellar sources with large magnitude errors when computing the positions of the locus points. Additionally, one wishes to eliminate as QSO candidates those sources that are far from the stellar locus only because their measurement error is large. As the algorithm currently stands, one will preferentially select as QSO candidates any objects with larger than average errors in their colors (since they will tend

to be outliers to the distribution), while missing some sources which are close to the stellar locus, but whose measurement errors are so small that their colors should be recognized as statistically separated from the stellar locus. Failing to account for these variable errors can result in spurious candidates at faint magnitudes (where magnitude errors are larger), and fewer candidates at brighter magnitudes (where magnitude errors are smaller).

Although it is possible to incorporate errors in the individual sources, it significantly complicates the algorithm. The first step is to choose a functional form for the underlying distribution of stars in the stellar locus. Previous searches for QSOs (with the notable exception of Gaidos, Magnier and Schechter 1993), have assumed that the stellar locus had no width at all. Therefore, it was not difficult to conceptualize a "distance from the stellar locus" or an "error distance to the stellar locus." We demonstrate in §3 that the stellar locus is *not* infinitesimally narrow. The cross sections that our algorithm measures along the locus are the result of both the intrinsic width of the stellar locus *and* the scatter produced by the magnitude errors in each measurement. To properly handle individual varying error measurements, one starts with a set of stellar sources with errors and produces from that the most likely distribution of "normal" stars. We assume that the stellar density when viewed in cross section at some point along the stellar locus is described by a two-dimensional Gaussian with arbitrary orientation. When choosing QSO candidates, one wishes to ask what the likelihood is that each source, with its individual errors, was drawn from the fit distribution of stars.

To accomplish this requires modifications at every stage in the algorithm. The stars associated with each locus point should be those which are closest in some "error distance" sense, instead of those that are closest in the Euclidean sense. Remember though, that we did not choose associated stars for each locus point entirely on their Euclidean distance because that did not lead to a stable solution. So, the procedure for choosing associated stars is complicated. When computing the ellipse fits at every locus point, the moments must be calculated as error-weighted moments. The criteria for adding new points between two existing locus points must be revisited in light of the fact that the size of the major axis no longer includes the typical errors in the data points. The centroids must be calculated with the new definition of associated stars, and also use the errors in each associated star to weight their contribution to the determination of the centroid. Outlier rejection must now use the likelihood that each source was drawn from our calculated underlying distribution rather than cutting out outliers based on their Euclidean distance from the stellar locus alone. Finally, the candidates must be chosen based on the likelihood that each source was drawn from the calculated distribution of normal stars.

Fortunately, good science can be accomplished without taking into account individual errors in each point. For instance, the locus fit currently implemented is applicable to a common case in which the width of the stellar locus is primarily due to photometric errors, and the photometric errors are approximately the same for all input sources (or at least are a function only of the color of the source). If there are enough data points, they can be grouped by photometric errors (usually corresponding to magnitudes) and the locus can be fit separately for each subset of the data with

similar errors. Or, the locus points can be fit to the whole distribution, and the ellipses can be fit separately for each subset of the data. Of course, the locus fit without errors is also valid for the case that the locus can be approximated with an elliptical (or cylindrical) cross section, and the photometric errors are small.

For cases in which the intrinsic width of the stellar locus is comparable to the photometric errors, the fits are useful but more difficult to interpret; cases where the input sources have very different photometric errors will require extensions to this algorithm before good color selection of QSOs can be accomplished. We hope to address the inclusion of photometric errors more extensively in future work.

3. The Locus of 7^{th} Magnitude Stars

We illustrate the use of parameterizing the stellar locus by applying these algorithms to *The Catalogue of WBVR Magnitudes of Northern Sky Bright Stars* (Kornilov et al. 1991). This catalog contains photometry for 13,586 stars, most with $V < 7.2$. Due to the bright apparent magnitude limit, the majority of the stars in the catalog are giants, particularly at the redder end of the locus.

We have plotted a color-color diagram of the WBVR data in Figure 2 (Color Plate XX). The W band is similar to, but slightly bluer than, Johnson U (Kornilov et al.). Stars of different spectral types have been plotted in different colors (some G stars are obscured by K stars in the plot). One sees the well-known bend at blue colors due to strong Balmer absorption in stellar classes A, B, and F. The effects of molecular blanketing are apparent in the colors of M stars. One striking feature in the plot is the visible effect of interstellar reddening, which moves stars up and to the right in the figure. The O stars, which are near their dusty birth places, are particularly reddened. From the locus of O stars, we determine the unit vector in the reddening direction for this filter system to be $\hat{A} = 0.583(W-B) + 0.622(B-V) + 0.523(V-R)$. We will assume that the reddening direction is not a strong function of stellar color.

Most of the stars were measured more than 4 times, with a standard error in each measurement of $\sigma_W = 0.010, \sigma_B = 0.007, \sigma_V = 0.006$, and $\sigma_R = 0.007$ (Kornilov et al.). Using the correlation coefficients, ρ_{WB}, ρ_{BV}, and ρ_{VR}, from Kornilov et al., we are able to calculate the standard error in the colors of one measurement from:

$$\sigma_{(W-B)} = \sqrt{\sigma_W^2 + \sigma_B^2 - 2\sigma_W\sigma_B\rho_{WB}} = 0.010$$

$$\sigma_{(B-V)} = \sqrt{\sigma_B^2 + \sigma_V^2 - 2\sigma_B\sigma_V\rho_{BV}} = 0.007$$

$$\sigma_{(V-R)} = \sqrt{\sigma_V^2 + \sigma_R^2 - 2\sigma_V\sigma_R\rho_{VR}} = 0.007$$

. Since the typical star has 4 measurements, the typical standard errors are $\sigma_{(W-B)} = 0.005, \sigma_{(B-V)} = 0.004$, and $\sigma_{(V-R)} = 0.004$. The phenomenal accuracy of this photometric data set will allow us to study the structure of the locus of 7^{th} magnitude stars.

From color-color plots, we choose endpoints of $\vec{r}_{start} = (W-B, B-V, V-R) = (-0.7, -0.2, -0.2)$ and $\vec{r}_{end} = (1.5, 1.7, 2.5)$. We choose $d_x = d_y = d_z = 0.4$ magnitudes to be the maximum distance from a locus point that an associated star could lie. The factor which determines the locus point spacing was chosen to be $N_{\sigma_{spacing}} = 3$, and the algorithm was allowed to iterate 100 times. After the preliminary set of locus points was derived, outlier rejection was applied with a 2 sigma rejection radius ($N_{\sigma_{reject}} = 2$) and 10 iterations. Table 1 shows the parameters produced; the black dots in Figure 2 show the positions of the locus points with respect to the sources. Note that the fit works well, despite the large degree of reddening in the data set.

In Figure 3 we have plotted cross sections in six places along the locus. Each plot in the figure contains sources with values of k that lie between the two indicated locus points. The plots show a)B−A stars, b) F stars, c) F−G stars, d) G−K stars, e) K stars, and f) M stars. Since we are plotting l vs. m in each plot, the major axes of the ellipse fits are always aligned with the horizontal axes. The larger ellipses in Figure 3 show the 3-sigma ellipse fits to the stellar locus cross sections. We notice from these plots that the major axes of the ellipse fits are *much* larger then the minor axes, and that the width of the stellar locus increases markedly for M stars. Also evident in the plots are a large number of sources outside the three sigma limits. Most of these sources are reddened stars from other parts of the stellar locus. For example, the sources on the right side of Figure 3a are reddened B stars. Figure 3b has reddened A and F stars on the right side, and reddened B stars on the left side. The only outliers that cannot be plausibly attributed to reddening are some of the M stars in Figure 3f. Many of these are variable stars or stars with unusual spectra.

The smaller ellipses at the base of the reddening vectors in Figure 3 show the approximate 3-sigma photometric measurement errors as calculated above, but taking into account the different photometric errors as a function of B−R, as tabulated in Kornilov et al.. Since correlation coefficients were given only for adjacent filters, we used the simplifying assumption that the correlation coefficients between different colors was zero for purposes of computing the error ellipses.

For convenience, the horizontal and vertical axes are labeled with the $\hat{l}$, $\hat{m}$ unit vectors, respectively, and projections of the color axes (of length 0.1 mag) are shown. One can see the increasing errors, especially in W−B, as the stars become redder. However, the widths (even in the thin direction) and orientation of the locus cross sections are not explained by the published catalog errors. One could imagine that there are unknown systematic errors in the catalog that would increase the errors by the factor of 2 or 3 required to explain the width in the thin direction. But whatever the cause of the thickness in the thin direction, one cannot attribute the width in the major axis direction or the broadening of the locus towards M stars to catalog errors.

One interesting feature of this data set is that there is no obvious separation of main sequence stars (medium line) and giants (dark line) in color-color-color space. To better understand this, we identified 3677 main sequence stars and 2871 giants in the data set. These are subsets of all

main sequence and giant stars in the catalog. Next, we computed the position of each data point in (l, m, k) space by comparing its position in color-color-color space with that of the closest locus point. Compared with the locus point, we know the data point's distance along the locus, distance along the thick direction of the locus, and distance along the thin direction of the locus. We compute each locus point's position along the locus by adding up the Euclidean distances between the locus points starting with the locus point in question and proceeding point by point back to the very first locus point at the beginning of the distribution.

Figure 4 presents a histogram showing the number of stars along the stellar locus as a function of position along the locus. Also shown are the distributions of main sequence stars and giants along the locus. The red end is predominately main sequence stars, and the blue end contains predominantly giants. In regions of the stellar locus which contain both main sequence and giant stars, we might expect to see some separation of the two distributions.

Figure 5 shows histograms across the stellar locus for stars in Figure 3d, a fairly straight section of the stellar locus. The histograms are fit to Gaussians of arbitrary height, but with centers given by the locus points and widths given by the ellipse fits. The bivariate Gaussian model for the density of stars in the locus cross section is a surprisingly good match to the data. As we previously mentioned, the sources in the tails of the histogram are mostly reddened stars, and were not used in determining the ellipse fits.

In Figure 5, we have also plotted the distributions of giants and main sequence stars. In these plots we see evidence for color separation of the main sequence and giant stars. In the direction of the minor axes, the center of the distribution of main sequence stars is displaced 0.008 ± 0.001 magnitudes to the right of the distribution of giant stars. The giant stars are marginally (0.0015 ± 0.0005 mag) displaced to the left of the distribution of all stars. In the direction of the major axis, the distribution of main sequence stars is consistent with the distribution of all stars. The distribution of giant stars is displaced by (0.006 ± 0.001 mag to the left of the other two distributions. This suggests a contribution to the stellar locus from at least one other population of stars. In both plots, the widths of the distributions of giant stars and main sequence stars is similar to the measured width of the distribution of all stars. The small separation between the locus of giant stars and main sequence stars in Figure 5 is not unexpected; the majority of the color separation between luminosity class III and V stars with the same spectral type is along the locus (Mihalas and Binney 1981).

The apparent intrinsic width of the stellar locus seen in Figures 3 and 5 cannot be explained by reddening. The arrows in the lower left corners of the plots in Figure 3 show the projection of the reddening vector (of length 0.1 magnitudes) onto the plotted plane. Although the reddening vector is close to being in line with the major axis of the ellipse (except Figure 3c), the length of the reddening vector does not correlate well with the width of the major axis, nor does it explain the Gaussian profile.

We can make a stronger case for ruling out reddening as a source of the observed width by

looking at the subset of the WBVR stars that have Galactic latitude $|b| > 60°$. Figure 6 shows the W-B and V-R colors of these 1218 stars. Notice that the large numbers of obviously reddened stars have disappeared. There are a few stragglers in a straight line from locus point 3 to locus point 9, and also above the locus of K stars. These probably represent the locus of supergiants.

We fit locus points to this high latitude stellar distribution using the input parameters $\vec{r}_{start} = (-0.7, -0.2, -0.2)$, $\vec{r}_{end} = (1.6, 1.7, 2.3)$, $d_x = d_y = d_z = 0.4$ magnitudes, $N_{\sigma_{spacing}} = 4$, and iterating 100 times. Outlier rejection was applied with $N_{\sigma_{reject}} = 2$ and 10 iterations. Table 2 shows the parameters produced, and Figure 7 shows the cross sections through the locus. The cross section through B-A stars was not calculated since there were too few stars in this region to obtain an accurate fit. Although we have successfully removed the obviously reddened stars, the measured width of the stellar locus has narrowed only slightly. The smaller catalog of input sources has, however, increased the errors both in the fit positions of the locus points and in the measurement of the major and minor axes at each point.

At nearly every place along the locus, when viewed in three dimensions, the Galaxy's stellar color locus appears thin, nearly as thin as the measurement errors in one cross-sectional dimension, but it appears many times thicker in the perpendicular dimension. This feature gives a 'ribbon-like' structure to the stellar locus when viewed in three dimensions. Since the ribbon twists along its length, one cannot see the flattened structure in one two-dimensional projection; rather, one must project different points along the locus at different angles to the WBVR axes to see the full effect.

We believe that this 'ribbon-like' structure to the stellar locus in the WBVR data set is a real effect, and have demonstrated that it is not due to measurement errors, reddening, mixtures of luminosity classes, or correlation between measurements in different filters. The high precision of the WBVR data allows one to see actual intrinsic width in the stellar color locus of our Galaxy. We believe that this is the first time such a measurement has been presented. All previous plots of the stellar locus have either had photometric errors too high to see intrinsic width (Koo, Kron, and Cudworth 1986) or they have plotted the data in a coordinate system which hid the intrinsic width (Gaidos, Magnier and Schechter 1993). Furthermore, stellar atmosphere models of stellar colors have concentrated on two-dimensional color-magnitude and color-color diagrams of the data (Bell and VandenBerg 1987), and they also have not previously noted a 'ribbon-like' structure. To see the intrinsic width of the locus, it is necessary to look at the data in more than two dimensions.

In the 'thick' direction, the full width at half maximum of the locus appears to be 0.07 magnitudes (see Figure 5). Presumably this intrinsic width is due to a combination of metallicity, gravity, and age differences between the stars.

4. Color selection of QSOs

In order to demonstrate the algorithm's ability to select QSOs in multicolor space, we are grateful to have been permitted pre-publication access to UJFN photographic ($\sigma \sim 0.1$ mag) and spectroscopic data in four fields from Koo and Kron 1988, Kron et al. 1991, Bershady et al. 1997. We selected only point sources (objects with stellar PSFs on the photographic plates as described in Kron et al. 1991) with magnitudes measured in all four filters, and with $18.0 < J < 21.5$. This selection resulted in 2604 point sources, of which 79 are confirmed QSOs as determined from broad emission lines in their spectra. An additional 257 objects have good spectra and are confirmed as stars, compact narrow emission line galaxies, or other non-quasar type objects. The objects have been studied for the last fifteen years, and QSOs have been identified by variability and proper motion (Trevese et al. 1994) as well as from colors.

We fit the stellar locus to this data set as before, using the input parameters $\vec{r}_{start} = (-0.2, 0.3, 0.2)$, $\vec{r}_{end} = (2.0, 2.0, 2.8)$, $N_{\sigma_{spacing}} = 3$, and $N_{iter} = 100$. Here, we have identified (x, y, z) with $(U-J, J-F, F-N)$. The outlier rejection used $N_{\sigma_{reject}} = 3$ and 10 iterations. Table 3 shows the parameters produced.

It is interesting to look for an intrinsic width to the stellar locus in this data set of faint point sources. The locus cross section at locus points 1-6 is roughly circularly symmetric (compare a_p, b_p in Table 3), with widths that are consistent with the photometric errors of the catalog, and are larger than the measured intrinsic width of the locus of 7^{th} magnitude stars. If there is a contribution to the measured cross section due to an intrinsic width of the stellar locus, it would require a more detailed analysis of this data set than we have attempted. One might be tempted to attribute the broadening of the stellar locus around points 7-10 to an intrinsic width. However, note that the major axis at the red end of the locus in is aligned almost exactly along the $U-J$ color axis. Using the parameters in Table 3 and the equations in §2.4, we find that the dot product of the major axis with the $U-J$ color axis, $\hat{l} \cdot \hat{x}$, is 0.97. Since we selected the objects based on a magnitude limit in the J band, it should be expected that the redder objects are systematically fainter in U, and therefore have greater errors.

Once the locus has been parameterized, it is simple to make a first order QSO selection. For each source, we identified the closest locus point; if the source was outside the $N_{\sigma_{select}} = 4$ sigma ellipse associated with that point, it was counted as a QSO candidate. Any source past the blue endpoint of the locus ($(\vec{r} - \vec{r}_{start}) \cdot \hat{k}_1 = 0$) was also flagged as a candidate. Sources past the red endpoint of the locus were compared with the ellipse associated with the reddest locus point to determine whether or not they were candidates. This process produced 184 candidates, of which 75 are confirmed QSOs and 24 are confirmed to be non-QSOs, with the remaining 85 candidates having no confirming spectra. Four confirmed QSOs were not selected as QSO candidates by this algorithm. The results of the selection are shown in Figure 8. Though Figure 8 plots only a two-dimensional projection (U−F vs. F−N) of the three-dimensional color space available, the selection of QSO candidates was done in three-space. Several candidate objects appear to be well

inside the stellar locus as projected in Figure 8. Yet, when viewed in three-space from a different direction, these objects are outside of the locus. The fact that some of these objects are confirmed QSOs demonstrates the usefulness of a multicolor selection algorithm which is able to look at more than two dimensions simultaneously.

We now examine completeness estimates for this algorithm. Completeness is defined to be the ratio of number of QSOs in the candidate sample to the total number of QSOs in the parent sample. A precise determination of completeness is difficult since it would require a spectrum of every object in a given field to a fixed magnitude limit. If the good spectra obtained in this sample were a fair subset of the data, then we would obtain a completeness of $75/79 = 95\%$. However, there is reason to believe that this calculation overestimates the completeness. Majewski et al. 1991 have done a more careful study of sample completeness, including a study of the variability and proper motion of all of the sample point sources. From their results, they estimate that there are 70 QSOs to $J = 22.5$ in the field SA57, which comprises one quarter of our data set. Of these, they estimate that 16, or 23%, are buried in the stellar locus, as they are not found by the color selection methods in Koo, Kron, and Cudworth 1986. Although we do not have the same color selection method or magnitude limit, it is likely that 5% is an underestimate of the fraction of QSOs that are buried in the locus of stars due to incompleteness in the spectroscopic sampling. The spectroscopic sample is probably less complete in the denser part of the stellar locus.

Selection efficiency, defined as the percentage of candidates selected by an algorithm which are actually QSOs, is easier to estimate reliably. Based on objects which have spectroscopic confirmations, our QSO selection efficiency is $75/(75+24) = 76\%$. Since 85 of our candidates have not been confirmed spectroscopically, we place upper and lower bounds on the efficiency of the algorithm at $(75+85)/(24+85+75) = 87\%$ and $75/(24+85+75) = 41\%$. Although the spectroscopic sample is probably not a fair sample for purposes of measuring completeness, it is less strongly biased for color selected QSOs. Therefore, 76% efficiency is a reasonable estimate for the performance of the algorithm on similar photometric catalogs. Note that the efficiency will change with sample depth and photometric accuracy.

In order to improve efficiency (at the expense of completeness), one may change the threshold, $N_{\sigma_{select}}$, which controls how far an object must be from the center of the stellar locus to be considered a candidate. Table 4 shows the results of varying this threshold from one to many sigma. Note that varying the cutoff threshold does not affect the efficiency for finding the lower redshift, bluer QSOs (which we call UVx QSOs). Therefore, we have included separate completeness figures for non-UVx QSOs in Table 4 as well. We define non-UVx QSO as one with colors that do not all lie blue-ward of the blue end clump in the stellar locus as seen in Figure 1. We chose $N_{\sigma_{select}} = 4$ as the best compromise between completeness and efficiency for this sample. Although this threshold was chosen *a posteriori*, we believe the results will not be much different even if the QSOs were not identified *a priori*; the efficiency and completeness are not a strong function of $N_{\sigma_{select}}$.

Table 5 shows the magnitude distribution for the 99 point-source objects with confirming spectra which were selected as QSO candidates by this algorithm. The sources are divided into stars, narrow emission line galaxies (NELGs), and QSOs. Note that the efficiency is relatively independent of magnitude for this data set. If the selection were strongly affected by the random errors in the magnitudes for each point, we would expect to pick up many more stars at fainter magnitudes than at brighter magnitudes. Since this is not the case, we must assume that the non-QSO candidates are due to systematic errors, local errors in the magnitudes (eg. plate defects), or authentically unusual colors of the objects. Therefore, it would not improve the results for this case to break the sample up into magnitude bins.

5. Comparison with Other Multicolor Selection Techniques

It is difficult to compare completeness and efficiency with other algorithms since the target selection goals and sample data vary widely. We have chosen to compare the results of our algorithm with the results from three other papers, all of which used at least four filters to select QSOs by color.

We first compare our results with those of Gaidos, Magnier and Schechter 1993. Their color selection technique is similar to ours in that they generate a set of locus points by iteratively moving them to the centroid of a local set of sources. Then, they select candidates based on a measure of the distance from the calculated locus points. Our method for generating the locus points is different in that it uses fewer locus points which are spaced according to the local width of the stellar locus, and defines associated stars between parallel planes. Both innovations tend to stabilize the final positions of the locus points (particularly at the wide, red end of the stellar locus). We also measure an elliptical fit to the locus cross section as a function of position along the locus, where Gaidos, Magnier and Schechter 1993 measured only one circular fit to the cross section. However, Gaidos, Magnier and Schechter 1993 did measure the error weighted distance from the stellar locus for each source when determining the locus points and when selecting candidates. For the case that the errors are much larger than the intrinsic width of the locus, this method will effectively eliminate stars that are in a volume of elliptical cross section, and the cross section will vary as a function of position along the locus. It will in fact do slightly better than our algorithm since the errors are treated individually for each source rather than as a bulk property of the data set. Currently, the only way to accommodate a significant variation of errors in each data point in our algorithm is to divide the data set up by its photometric errors, as described in section 2.10, and select QSOs from each subset individually. The efficiency of the Gaidos, Magnier and Schechter 1993 algorithm is quoted at 80%, based on BVRI multicolor selection. However, they have not obtained confirming spectra for their data so their efficiency is only an estimate.

One of the largest advantages our algorithm has over that of Gaidos, Magnier and Schechter 1993 is our ability to study the locus in cross section and along its length. This makes it easier for us to measure the stellar locus intrinsic width, and to study data sets that are not dominated by

measurement error. Gaidos, Magnier and Schechter 1993 were forced to resort to a Monte Carlo simulation to study the structure of the stellar locus in their data. The cross section projections of Fig. 5 in Gaidos, Magnier and Schechter 1993 are not exactly perpendicular to the locus direction as noted in their text, and thus their apparent ellipses are an artifact of a projection (Gaidos 1997). However, these authors, even with a data set with errors of about 0.1 mag, were able to place an upper limit on the intrinsic width of the stellar locus of 0.1 mag. Their adopted intrinsic width, 0.075 mag, is surprisingly close to the width we determined from the locus of 7^{th} magnitude stars in §3.

Next, we compare the performance of our algorithm with that of Kron et al. 1991, Bershady 1996. We used a subset of the same catalog they used to test their algorithm. Like our case, the algorithm described in their paper was not the one used to select the candidates for follow-up spectra, and they quote efficiency and completeness from an algorithm whose input parameters were determined *a posteriori*. Their locus points are also generated somewhat differently. First, a locus point is placed in the position of each stellar source. Locus points that are in the lowest density of multicolor space are removed so that they do not disrupt the convergence of the locus points to the center of the stellar locus. Then, the locus points are iteratively moved in $(U-J, J-F, F-N)$ space to the centroid of the local set of locus points. This has the disadvantage that locus points tend to be pulled out of regions of lower stellar density, which can create a blue clump and a red clump of locus points. The locus points must be re-ordered so that the tangent to the curve down the center of the stellar locus can be calculated. Candidates were chosen as those points which exceeded thresholds in the 5^{th} nearest neighbor distance as well as error weighted distance from the locus. Note that in this algorithm, the threshold for distance from the locus is adjusted for width variations along the locus only if they are due to photometric errors; the selection does not account for any intrinsic spread in the locus. The red (M-dwarf) end ($J-F>1$ and $F-N>1$) of the locus was excluded.

Kron et al. 1991 obtained 89% completeness and 69% efficiency for their sample of 130 spectroscopically confirmed QSOs. These numbers were calculated in exactly the same way we calculated our 95% completeness and 76% efficiency. The same caveats for our calculation of the efficiency, and especially the completeness, also apply to these numbers. The comparison is not as meaningful as one would have hoped to obtain for two reasons. First, the source catalogs used in these two cases were not exactly the same. They used magnitude cut-offs one magnitude fainter than ours (J < 22.5 and F < 21.5), and also excluded objects at the red end of color space, with $J-F>1$ and $F-N>1$. Second, the spectroscopic sample is not complete enough, and therefore the results are not accurate enough, to make detailed comparisons. Their results and ours are the same within the accuracy of this test.

Last, we compare our results with those of Warren et al. 1991a, and Warren et al. 1991b. The essential features of this algorithm, which uses 10^{th} nearest neighbor criterion for selecting QSOs, were discussed in §1. Their data has photometry in six optical filters, and is complete for $16 < m_{or} < 20$. In addition to the density criterion in six-dimensional magnitude space, they

exclude UVx sources (although their criteria are somewhat different than ours) in order to focus on high-redshift QSOs. They also exclude sources on the red end of the stellar locus. Of the 19 previously discovered QSOs with $z \geq 2.2$ in their data set, they found 15, yielding a completeness of 79%. However, the completeness drops precipitously for objects with $m_{or} > 18.5$, since their algorithm preferentially selects brighter objects. The best characterization of their efficiency is 43% for QSOs with $z \geq 2.2$. Since these results do not include UVx QSOs, we should compare with our results for non-UVx candidates. From Table 4, we find $75 - 66 = 9$ non-UVx QSOs and $24 - 14 = 10$ non-UVx non-QSOs in a sample with $79 - 66 = 13$ known QSOs. This gives us a completeness of $9/13 = 69\%$ and an efficiency of $9/19 = 47\%$. Again, their results are comparable to the results obtained by our algorithm.

6. Discussion and Results

We have presented an algorithm that successfully selects QSOs from multicolor photometric data. We have seen that it is also useful for studying the structure of the stellar locus, given a data set with high photometric accuracy. The algorithm has also been used to study the photometric errors in stellar data sets. By fitting the locus points separately to subsets of a data set and comparing the results, it is possible to detect calibration errors or real differences in the position of the locus from place to place in the sky or for different magnitude ranges.

The main result of this work is the formal definition of an efficient and flexible algorithm for parameterizing a locus in n-dimensional space. When applied to a catalog of Galactic stars with precise photometry, it produced the quantitative result, directly apparent from examining column 9 of Table 1, that the stellar locus for $V < 7.2$ stars has an intrinsic width of FWHM$\sim$ 0.07 magnitudes along much of its length in WBVR space. The width is signficantly larger at the red end of the locus. We have shown that the width is not due to photometric errors, reddening, or differences in luminosity class.

When used to assist in multicolor selection of QSOs, the algorithm parameterizes Galactic stars so they can be removed from consideration. The algorithm allows any spread in the locus to be taken into consideration in a systematic fashion, so that, for instance, the threshold for outlier detection is further from the locus at the red M-dwarf end of the locus than at the blue end. As applied to a sample of four color data, the algorithm selects 95% of the spectroscopically confirmed QSOs with an efficiency of approximately 76%. Accurate comparisons of the results of our algorithm with others would require running all algorithms on the same, very completely studied, set of input data. The comparisons we were able to do show that the results of our algorithm are at least as good as those of other multicolor selection techniques. Our results were obtained with few adjustable parameters, a more simply defined algorithm, and no special refinements.

The algorithm is quite general; it does not currently assume anything about the data set other than that it contains points which are distributed in a more or less one-dimensional subset of

n-dimensional space. As written, the algorithm could digest weather data as easily as astronomical data. The data can then be transformed into a coordinate system that includes distance along the ridge and distance along the principal axes, where the directions of the principal axes can rotate around the ridge-line from one point on the ridge to the next. As we have shown, this new coordinate system can allow us to view the data in physically important ways.

We thank Rich Kron, Jeff Munn, Matt Bershady, John Smetanka, David Koo, and Steve Majewski for allowing access to a unique photometric and spectroscopic data set of stellar objects, and especially Rich Kron and Jeff Munn for assisting us in its use. Matt Bershady provided useful discussions regarding his QSO candidate selection algorithm in Kron et al. 1991. We also extend thanks to Victor Kornilov for providing us with a description of the WBVR catalog (in English) and assisting us with its interpretation. We acknowledge Lee Newberg for statistical assistance and for useful comments on the manuscript. We also thank Don Petravick for his assistance in locating information on nearest-neighbor algorithms. We are grateful to Fermi National Accelerator Laboratory for support of this work.

Table 1. Parameterization of the locus of 7^{th} magnitude stars in WBVR

	k	N_p	x_p	y_p	z_p	k_{x_p}	k_{y_p}	k_{z_p}	a_p	b_p	θ_p
1	0.000	383	−0.469	−0.073	−0.034	0.991	0.124	0.060	0.038	0.012	−0.828
2	0.268	427	−0.203	−0.040	−0.018	0.982	0.171	0.085	0.037	0.012	−0.799
3	0.484	659	0.005	0.010	0.008	0.876	0.399	0.271	0.026	0.013	−0.937
4	0.643	891	0.111	0.104	0.080	0.313	0.732	0.606	0.043	0.013	−1.408
5	0.802	779	0.097	0.223	0.184	−0.332	0.696	0.636	0.055	0.013	1.270
6	0.967	610	0.007	0.322	0.280	−0.624	0.574	0.530	0.056	0.010	1.102
7	1.149	734	−0.119	0.421	0.367	−0.366	0.728	0.579	0.048	0.009	1.352
8	1.294	610	−0.102	0.538	0.451	0.409	0.755	0.513	0.053	0.009	−1.194
9	1.473	264	0.009	0.656	0.527	0.736	0.563	0.377	0.040	0.009	−0.908
10	1.728	166	0.215	0.781	0.613	0.789	0.521	0.327	0.056	0.019	−1.006
11	1.926	239	0.366	0.892	0.675	0.807	0.525	0.271	0.034	0.010	−0.852
12	2.052	328	0.475	0.950	0.701	0.879	0.433	0.201	0.027	0.010	−0.731
13	2.152	365	0.564	0.990	0.720	0.884	0.410	0.225	0.027	0.011	−0.710
14	2.246	372	0.646	1.029	0.744	0.868	0.421	0.263	0.029	0.010	−0.653
15	2.337	329	0.725	1.068	0.769	0.887	0.397	0.234	0.028	0.011	−0.738
16	2.447	273	0.825	1.109	0.792	0.896	0.384	0.223	0.029	0.010	−0.741
17	2.545	275	0.912	1.148	0.815	0.865	0.417	0.278	0.033	0.013	−0.804
18	2.661	254	1.009	1.198	0.851	0.865	0.411	0.288	0.041	0.014	−0.758
19	2.786	240	1.120	1.247	0.885	0.864	0.414	0.288	0.047	0.014	−0.729
20	2.940	228	1.250	1.314	0.931	0.851	0.419	0.316	0.052	0.014	−0.735
21	3.108	179	1.393	1.382	0.987	0.836	0.419	0.355	0.031	0.013	−0.758
22	3.278	194	1.533	1.455	1.051	0.818	0.429	0.384	0.034	0.011	−0.845
23	3.459	194	1.680	1.532	1.121	0.794	0.409	0.450	0.027	0.015	−1.181
24	3.646	207	1.824	1.605	1.216	0.643	0.396	0.656	0.027	0.019	−1.051
25	3.827	217	1.912	1.675	1.358	0.187	0.253	0.949	0.052	0.019	−0.178
26	4.008	147	1.887	1.691	1.537	−0.367	−0.036	0.930	0.096	0.021	0.581
27	4.233	95	1.767	1.661	1.725	−0.593	−0.142	0.792	0.080	0.026	0.396
28	4.519	62	1.585	1.618	1.940	−0.639	−0.149	0.755	0.069	0.030	0.320

Note. — This table contains one line for each of the 28 locus points along the stellar locus. For each locus point we record: column 1 (k) - the distance in magnitudes along the stellar locus, starting from the first locus point; column 2 (N_p) - the number of sources associated with the locus point (not including the outliers which were rejected); columns 3, 4, 5 (x_p, y_p, z_p) - the $W-B, B-V, V-R$ position of the locus point; columns 6, 7, 8 ($k_{x_p}, k_{y_p}, k_{z_p}$) - the components of the unit vector $\hat{k_p}$ along the locus; and columns 9, 10, 11 (a_p, b_p, θ_p) - the major axis, minor axis, and position angle of the ellipse fit to the cross section of the stellar locus.

Table 2. Parameterization of the locus of WBVR stars at high galactic latitude

	k	N_p	x_p	y_p	z_p	k_{x_p}	k_{y_p}	k_{z_p}	a_p	b_p	θ_p
1	0.000	88	0.055	0.034	0.020	0.249	0.747	0.617	0.049	0.010	−1.330
2	0.180	83	0.100	0.168	0.131	−0.050	0.749	0.661	0.042	0.011	1.538
3	0.335	74	0.040	0.271	0.229	−0.529	0.611	0.589	0.045	0.009	1.167
4	0.505	61	−0.070	0.365	0.320	−0.568	0.610	0.552	0.052	0.008	1.155
5	0.635	93	−0.130	0.453	0.394	−0.013	0.805	0.593	0.056	0.010	−1.518
6	0.786	66	−0.073	0.570	0.472	0.581	0.675	0.455	0.050	0.007	−1.086
7	1.020	25	0.090	0.709	0.567	0.738	0.565	0.369	0.044	0.013	−0.828
8	1.309	25	0.312	0.865	0.665	0.780	0.538	0.321	0.035	0.013	−0.613
9	1.438	35	0.416	0.933	0.700	0.893	0.416	0.172	0.029	0.008	−0.787
10	1.555	34	0.528	0.966	0.706	0.924	0.349	0.157	0.027	0.005	−0.549
11	1.642	30	0.602	1.004	0.732	0.819	0.476	0.321	0.022	0.009	−0.699
12	1.709	29	0.654	1.039	0.755	0.874	0.421	0.244	0.028	0.007	−0.627
13	1.782	26	0.722	1.062	0.766	0.924	0.343	0.169	0.021	0.008	−0.561
14	1.838	17	0.773	1.083	0.777	0.885	0.410	0.220	0.020	0.004	−0.783
15	1.925	34	0.849	1.120	0.797	0.902	0.372	0.221	0.029	0.006	−0.518
16	2.013	31	0.930	1.148	0.816	0.891	0.388	0.233	0.028	0.012	−0.708
17	2.109	25	1.012	1.191	0.840	0.842	0.441	0.310	0.044	0.013	−0.724
18	2.230	25	1.112	1.243	0.883	0.882	0.396	0.257	0.039	0.009	−0.921
19	2.389	22	1.258	1.302	0.911	0.872	0.395	0.290	0.040	0.009	−0.751
20	2.550	19	1.390	1.369	0.975	0.797	0.435	0.418	0.014	0.012	−0.978
21	2.718	30	1.520	1.445	1.049	0.812	0.426	0.399	0.037	0.016	−0.632
22	2.920	19	1.690	1.526	1.122	0.775	0.396	0.492	0.020	0.010	−0.576
23	3.123	31	1.830	1.603	1.246	0.427	0.333	0.841	0.037	0.019	−0.129
24	3.361	26	1.868	1.665	1.473	−0.125	0.096	0.988	0.057	0.017	1.396
25	3.649	8	1.768	1.651	1.742	−0.510	−0.121	0.852	0.090	0.014	0.445
26	3.922	13	1.587	1.599	1.941	−0.658	−0.194	0.727	0.089	0.027	0.179

Note. — The column descriptions are the same as for Table 1.

Table 3. Parameterization of the UJFN data

	k	N_p	x_p	y_p	z_p	k_{x_p}	k_{y_p}	k_{z_p}	a_p	b_p	θ_p
1	0.000	479	0.027	0.707	0.324	0.760	0.580	0.293	0.136	0.095	0.055
2	0.379	290	0.315	0.927	0.435	0.798	0.523	0.299	0.136	0.099	0.297
3	0.770	268	0.640	1.108	0.554	0.824	0.469	0.318	0.123	0.081	0.325
4	1.234	245	1.019	1.327	0.707	0.759	0.510	0.405	0.148	0.101	0.015
5	1.691	291	1.335	1.576	0.925	0.577	0.518	0.632	0.166	0.115	-0.262
6	2.100	319	1.508	1.767	1.243	0.272	0.288	0.918	0.183	0.128	-0.564
7	2.498	269	1.547	1.799	1.638	0.111	0.053	0.992	0.161	0.122	-0.088
8	2.952	211	1.602	1.812	2.089	0.123	0.027	0.992	0.188	0.115	0.026

Note. — The column descriptions are the same as for Table 1.

Table 4. Completeness and Efficiency of the QSO selection algorithm

$N_{\sigma_{select}}$	QSO cand. (out of 79)	completeness (all QSOs)	completeness (non-UVx QSOs)	non-QSO cand. (out of 257)	unknown cand.	total cand.	efficiency (obj. with spectra)
1	79	1.00	1.00	193	1368	1640	0.29
2	79	1.00	1.00	70	343	492	0.53
3	76	0.96	0.77	28	118	222	0.52
4	75	0.95	0.69	24	85	184	0.76
5	74	0.94	0.62	20	66	160	0.79
6	73	0.92	0.59	16	57	146	0.82
10	70	0.89	0.31	14	46	130	0.83
100	66	0.84	0.00	14	46	126	0.83

Table 5. Magnitude Distribution for Candidates of Known Type

J	# of stars	# of NELGs	# of QSOs
18.0 – 18.5	2	0	0
18.5 – 19.0	3	0	3
19.0 – 19.5	1	0	5
19.5 – 20.0	0	0	9
20.0 – 20.5	0	1	11
20.5 – 21.0	3	5	24
21.0 – 21.5	3	6	23

REFERENCES

Bell, R. A. and VandenBerg, D. A. 1987, ApJ 63, 335

Bershady, M. 1996, personal communication

Bershady, M. A., Munn, J. A., Majewski, S., Kron, R. A., Koo, D. C., and Smetanka, J. J. 1997, in preparation

Braccesi, A., Lynds, R., and Sandage, A. 1968, ApJ, 152, L105

Braccesi, A., Formiggini, L., and Gandolfi, E. 1970, A&A, 5, 264

Fukunaga, K., and Narendra, P. (1975), IEEE Transaction on Computers, 24, 750

Gaidos, E. J. 1997, personal communication

Gaidos, E. J., Magnier, E. A., and Schechter, P. L. 1993, PASP, 105, 1294

Green, R. F., Schmidt, M., and Liebert, J. 1986, ApJS, 61, 305

Gunn, J. E. 1995, BAAS, 186, #44.05

Hall, P. B., Osmer, P. S., Green, R. F., Porter, A. C., and Warren, S. J. 1996, ApJ, 462, 614

Hazard, C. 1979, Active Galactic Nuclei, C. Hazard and S. Mitton, Cambridge University Press, Cambridge, 1

Janes, K. and Lynds, R. 1969, ApJ, 155, L47

Koo, D. C., Kron, R. G., and Cudworth, K. M. 1986, PASP, 98, 285

Koo, D. C., and Kron, R. G. 1988, ApJ, 325, 92

Koo, D. C., and Kron, R. G. 1982, A&A, 105, 107

Kornilov, V., Mironov, A., and Zakharov, A. 1996, Baltic Astronomy, 5, 379

Kornilov, V., and Mironov, A. 1994, Science with Astronomical Near-Infrared Sky Surveys, N. Epchtein, A. Omont, B. Burton, and P. Persi, Kluwer Academic Publishers, 83

Kornilov, V. G., Volkov, I. M., Zakharov, A. I., Kozyreva, V. S., Kornilova, L. N., Krutjakov, A. N., Krylov, A. V., Kusakin, A. V., Leontiev, S. E., Wironov, A. V., Moshkaliov, V. G., Pogrosheva, T., Sementsov, V. N., and Khaliullia, Kh. F. 1991, Trudy Gosud. Astron. Sternberga, 63, 4

Kron, R. A., Bershady, M. A., Munn, J. A., Majewski, S., and Koo, D. C. 1991, The space distribution of quasars, D. Crampton, Astron. Soc. Pac. Conf. Ser., 21, 32

Majewski, S. R., Munn, J. A., Kron, R. G., Bershady, M. A., Smetanka, J. J., Koo, D. C. 1991, The Space Distribution of Quasars, D. Crampton, Astron. Soc. Pac. Conf. Ser., 21, 55

Mihalas, D., and Binney, J. 1981, Galactic Astronomy: Structure and Kinematics, W. H. Freeman and Co., New York, p.108

Porter, A., Campin, M., Ogle, P., Maraziti, D., Green, R., and Osmer, P. 1991, The Space Distribution of Quasars, D. Crampton, Astron. Soc. Pac. Conf. Ser., 21, 88

Trevese, D., Kron, R. G., Majewski, S. R., Bershady, M. A., and Koo, D. C. 1994, ApJ, 433, 494

Warren, S. J., Hewett, P. C., Irwin, M. J., and Osmer, P. S. 1991, ApJS, 76, 1

Warren, S. J., Hewett, P. C., and Osmer, P. S. 1991, ApJS, 76, 23

Zhan, Y., Koo, D. C, and Kron, R. G. 1989, PASP, 101, 631

Fig. 1.— Algorithm Demonstration. We show here the progress of the algorithm with each iteration. The algorithm starts out (0 iterations) by putting a locus point at the position of each endpoint. For purposes of fitting locus points, the positions of the endpoints need not be very accurate. During the first iteration, a new locus point is added at the midpoint between the first two locus points, the ellipse fits are calculated, and then the three locus points are moved to the centroids of their associated stars. Since the ellipse fits were calculated before the locus points were moved, the calculated ellipse widths are large. Therefore, on the second iteration the locus points are closer together than 3 times the larger sigmas of the ellipses, so no new locus points are added. The sigmas are now re-calculated, and the three locus points are again moved to the centroid of their associated stars. Note that the set of associated stars for each point in the second iteration is different from the set of associated stars in the first iteration. During each iteration up to iteration 100, the same three steps: add points, calculate sigmas, move points - are repeated. The last plot shows the positions of the locus points after sigma rejection has been run. Because the outliers of this distribution are fairly symmetric, the locus points hardly moved. However, outlier rejection decreased the size of the major axes of the ellipse fits at each locus point by 17% on average.

Fig. 2.— The Catalog of WBVR Magnitudes of Northern Sky Bright Stars. We show the catalog stars, color coded by spectroscopic type. Due to the bright apparent magnitude limit of the catalog (about 7^{th} magnitude), the majority of the stars are giants, including almost all of the stars on the red end of the locus. Also shown are 28 locus points that were fit to the stellar locus. Redder stars were plotted over the bluer stars, obscuring some of the bluer stars. In particular, G stars run from about point 8 to point 15. Reddening, particularly of hotter, bluer stars is immediately apparent as a smearing out of the data towards the upper right with a slope of ≈ 0.9.

Fig. 3.— Cross Sections Through the Locus of 7^{th} Magnitude Stars. Cross sections are shown for six places along the locus. Refer to Figure 2 for the positions of the locus points. Each cross section includes the 3σ ellipse fit, a smaller ellipse showing the estimated 3σ measurement errors in each point, the projected directions of the three color axes, and the projection of the reddening vector as determined from the locus of O stars. By construction, the fit locus points go through (0,0) on each plot, and the major axis of the ellipse fit is along the x-axis.

Fig. 4.— Distribution of Stars Along the Stellar Locus of 7^{th} Magnitude Stars. The density of stars in the stellar locus is plotted as a function of the distance along the locus (in magnitudes) with 0 corresponding to the position of locus point 1 (light line). Also shown are the distribution of 3677 main sequence stars (medium line), and the distribution of 2871 giant stars (dark line) within the sample. Neither the set of main sequence stars nor the set of giant stars contains all of the WBVR stars in its luminosity class.

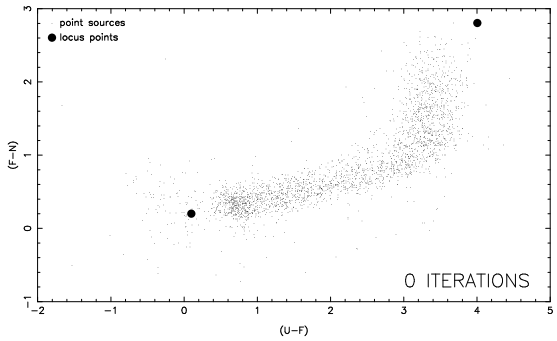
Fig. 5.— Distribution of Stars in the Stellar Locus Cross Section. Figure 3d shows the cross section of the locus of 7^{th} magnitude stars between locus points 12 and 16 (G and K stars). We show two separate plots through that cross section - one along the minor axis of the fit ellipse, and one along the major axis of the fit ellipse. The three histograms in each plot show all stars (light line), giant

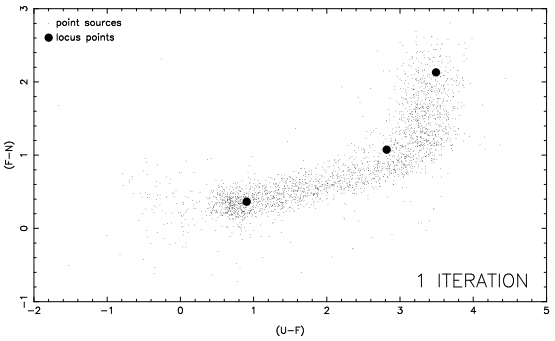
stars (medium line), and main sequence stars (dark line). Again, the sets of giant stars and main sequence stars are not complete. Also shown is a Gaussian centered at zero with sigma given by the algorithm's fit to the width of the locus. The amplitude of the Gaussian was adjusted to fit the histogram. The stellar density of the locus cross sections in color-color-color space are fairly good fits to a Gaussian, but with larger tails. Most of the stars in the tails are probably reddened stars.

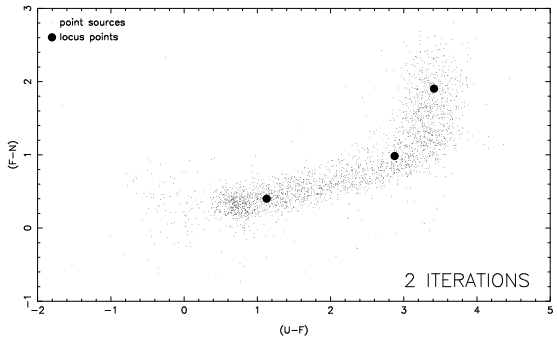
Fig. 6.— WBVR Stars at High Galactic Latitude. The small dots are the stars from Figure 2 with $|b| > 60°$ Galactic latitude. The large dots are 26 locus points fit to these stars. The positions of the locus points and their associated parameters are tabulated in Table 2.

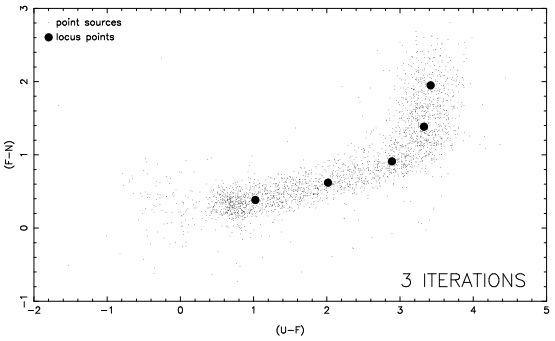
Fig. 7.— Cross Sections Through WBVR Stars at High Galactic Latitude. Cross sections are shown for five places along the locus in Figure 6. The cross sections are from similar, but not identical, places to the cross sections in Figure 3. The cross section through the B-A stars was omitted since there were not enough high latitude stars in that region to make an accurate fit.

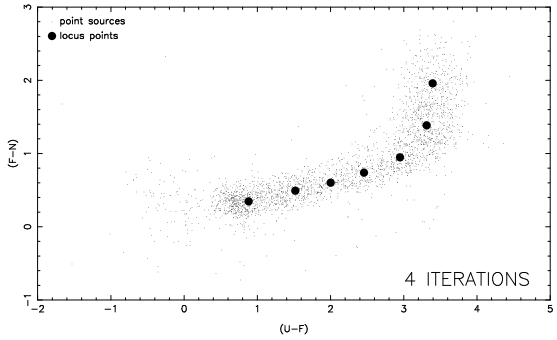
Fig. 8.— Efficiency for Finding Quasars. We tested the ability of our algorithm to find QSOs using a well-studied sample of 2604 point sources with $18.0 < J < 21.5$ from four sets of UJFN photographic plates. High quality spectra have been obtained for 336 of these sources, including over half of our candidate QSOs. The sources for which high quality spectra have not been obtained are shown as dots; broad line emission sources (QSOs) have been plotted as filled squares; and sources without broad-line emission have been plotted as asterisks. The large filled circles show the positions of the locus points that have been fit to the locus of stars. We selected as candidate QSOs (open circles) those objects that were outside of a four sigma ellipse fit to the cross section of the stellar locus at each locus point. Of the 184 candidates we selected, 75 of the sources are known QSOs, 24 are known to be stars or narrow emission line galaxies, and 85 are spectroscopically unidentified.

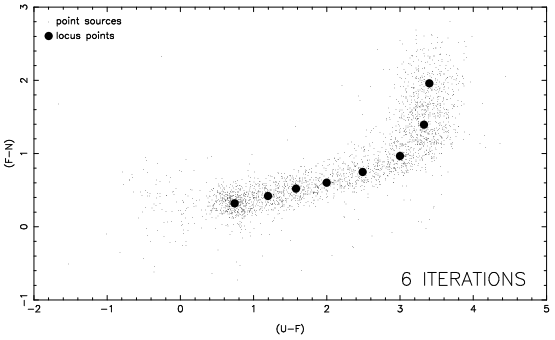

point sources
locus points
0 ITERATIONS
(F−N)
(U−F)

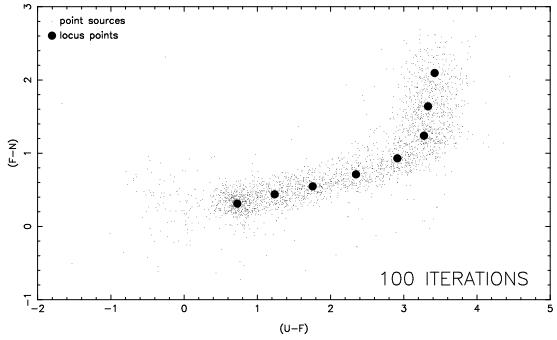

point sources
locus points
1 ITERATION
(F−N)
(U−F)

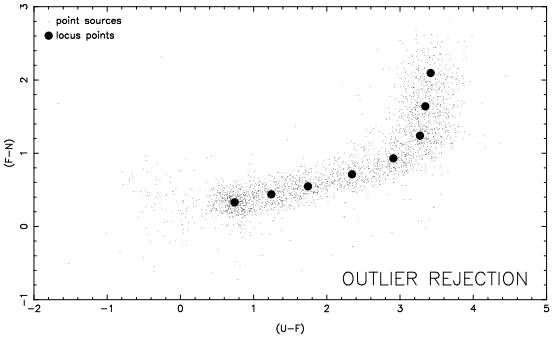

point sources
locus points
2 ITERATIONS
(F−N)
(U−F)

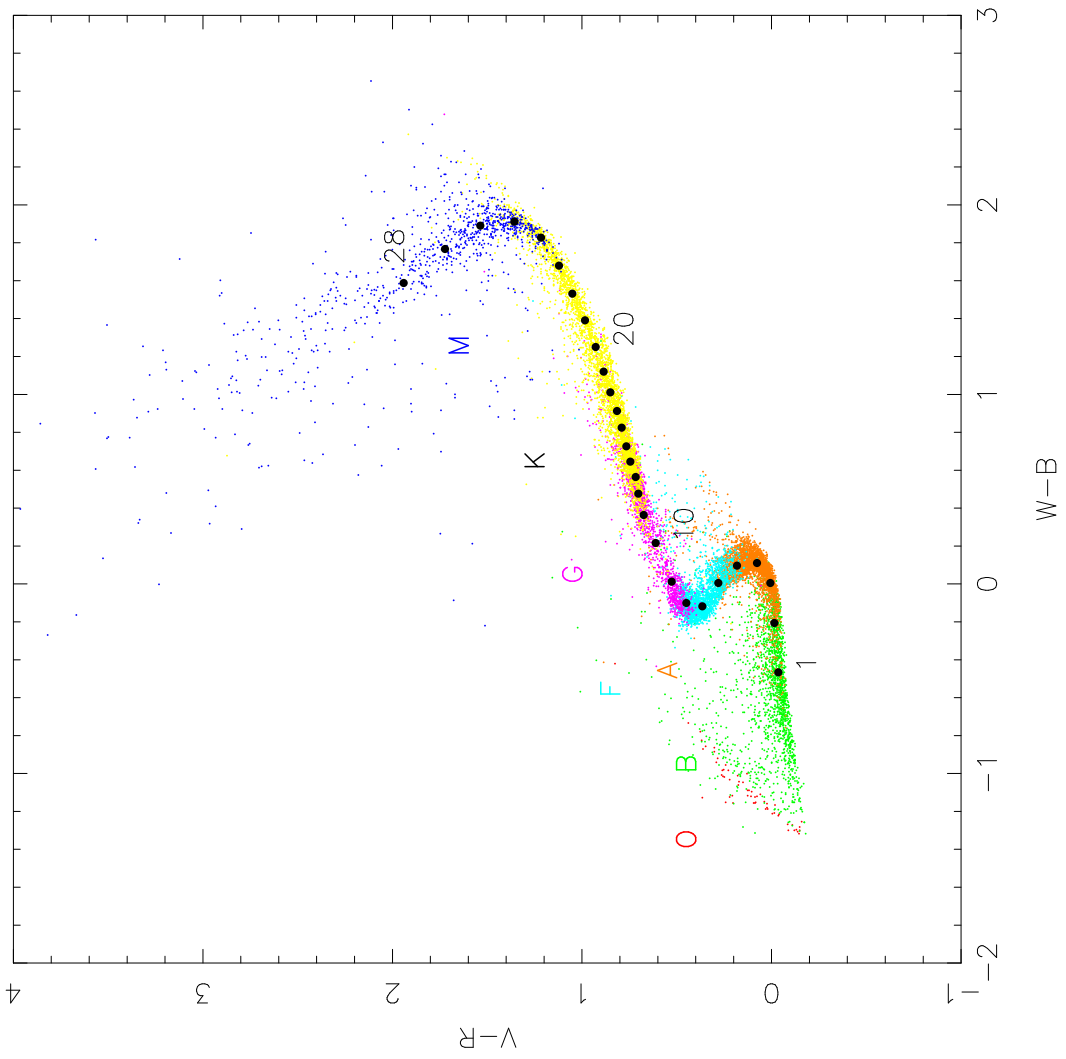

point sources
locus points
3 ITERATIONS
(F−N)
(U−F)

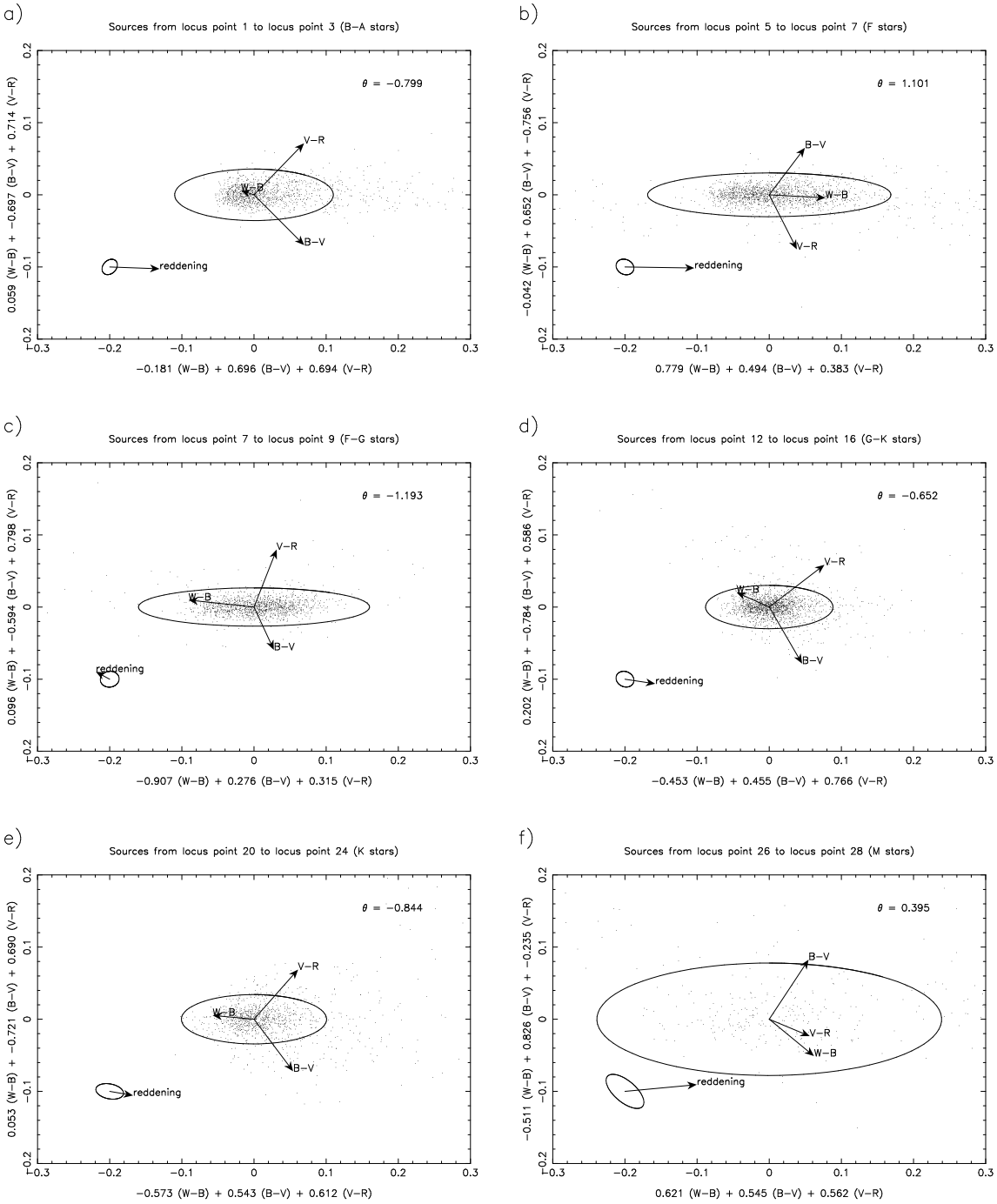

point sources
locus points
4 ITERATIONS
(F−N)
(U−F)

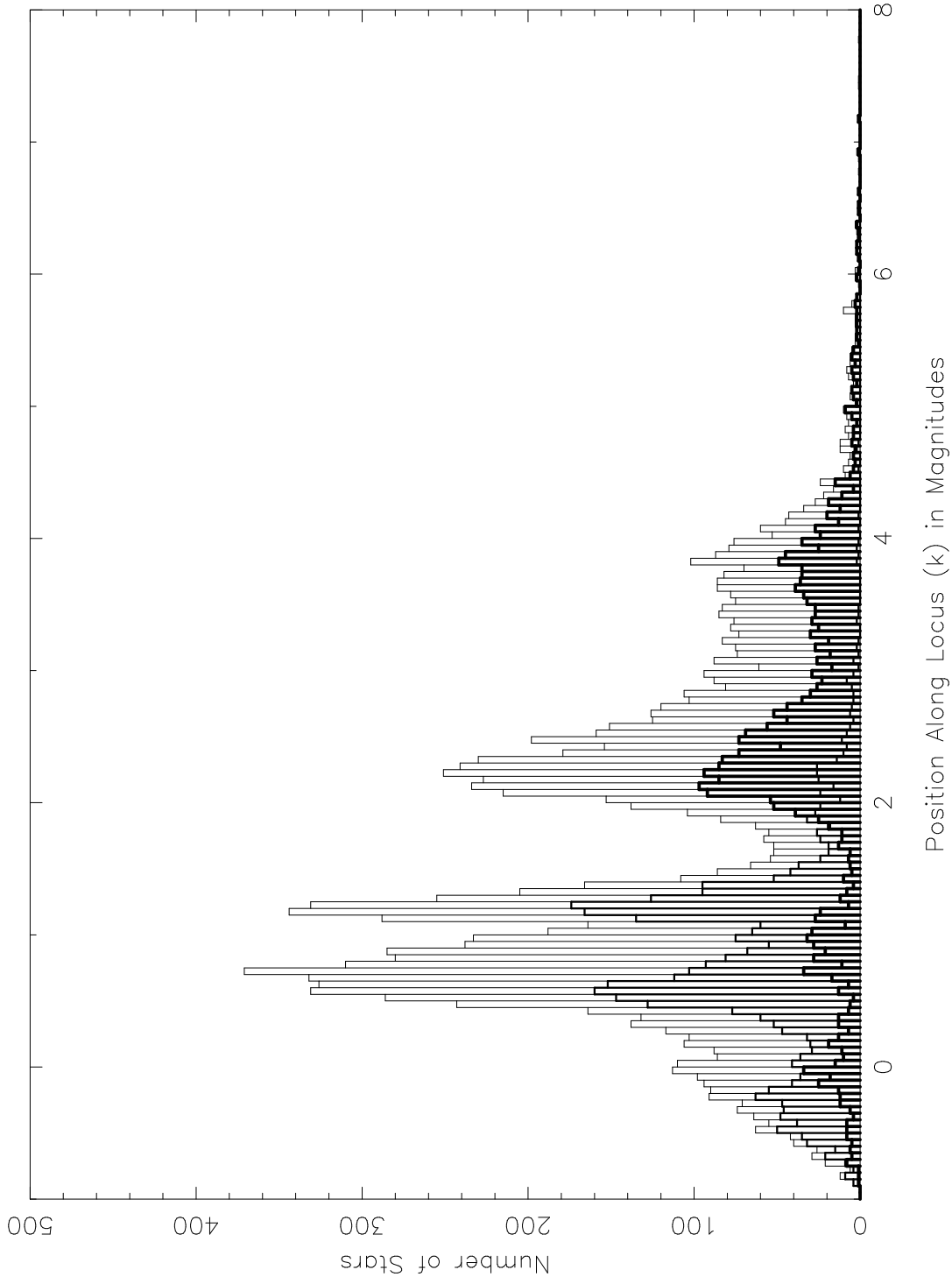

point sources
locus points
6 ITERATIONS
(F−N)
(U−F)

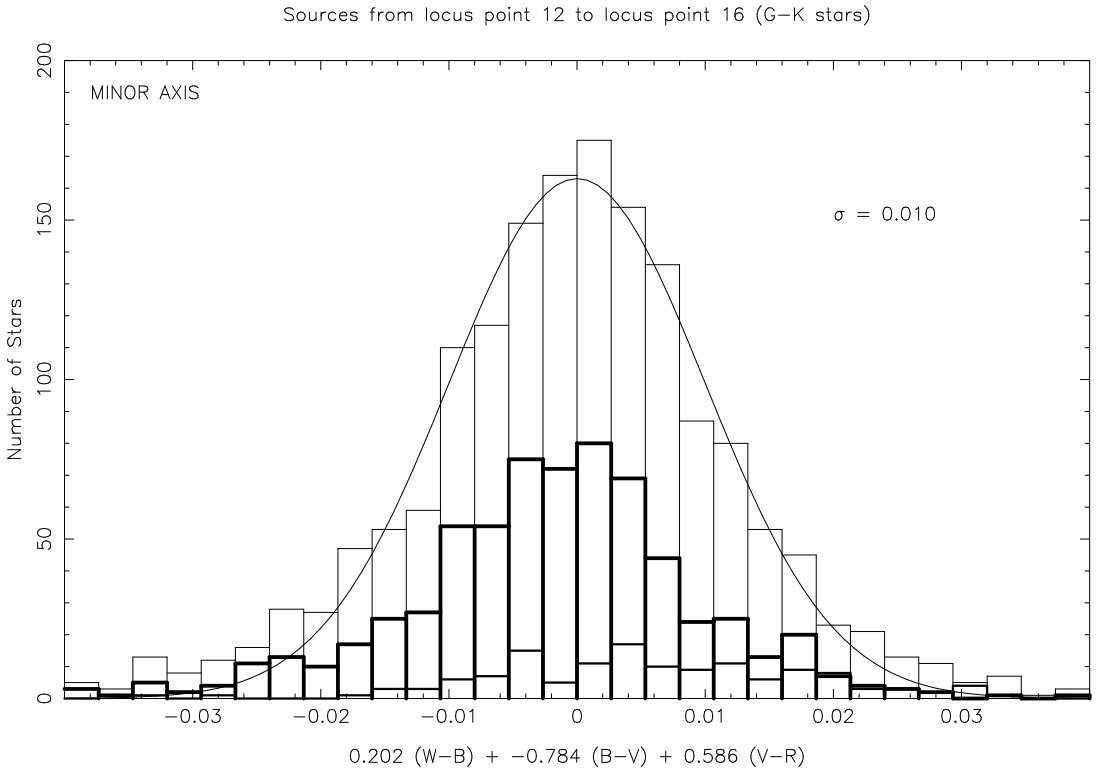

point sources
locus points
100 ITERATIONS
(F−N)
(U−F)

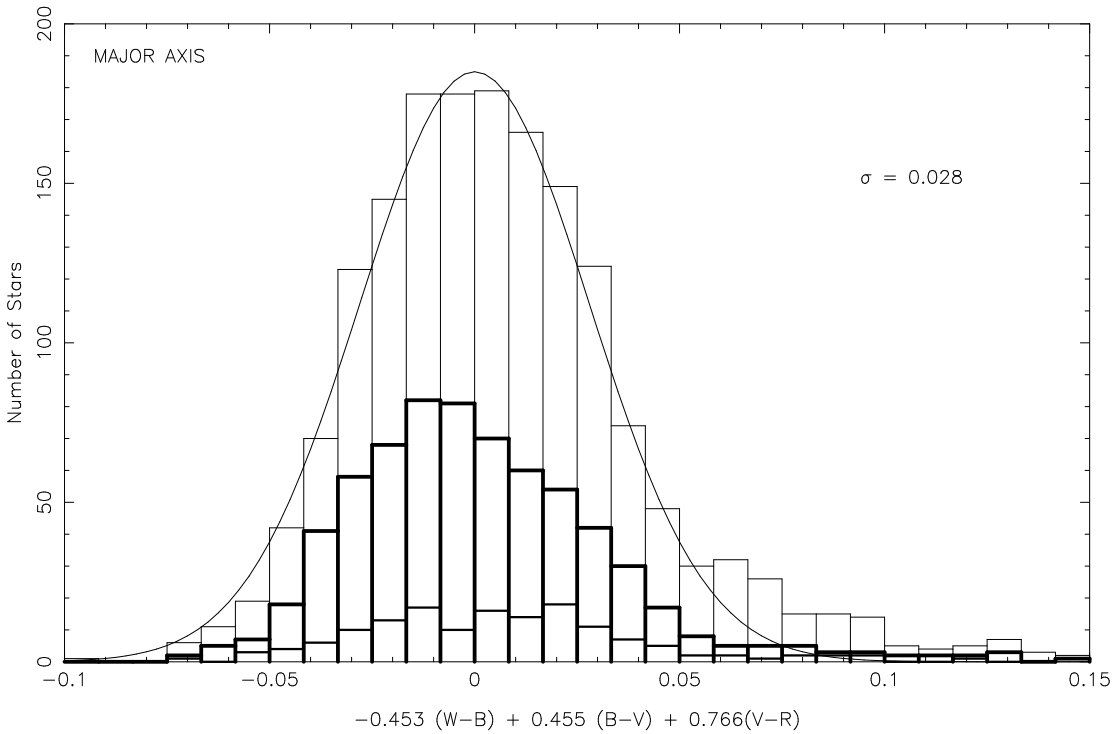

point sources
locus points
OUTLIER REJECTION
(F−N)
(U−F)

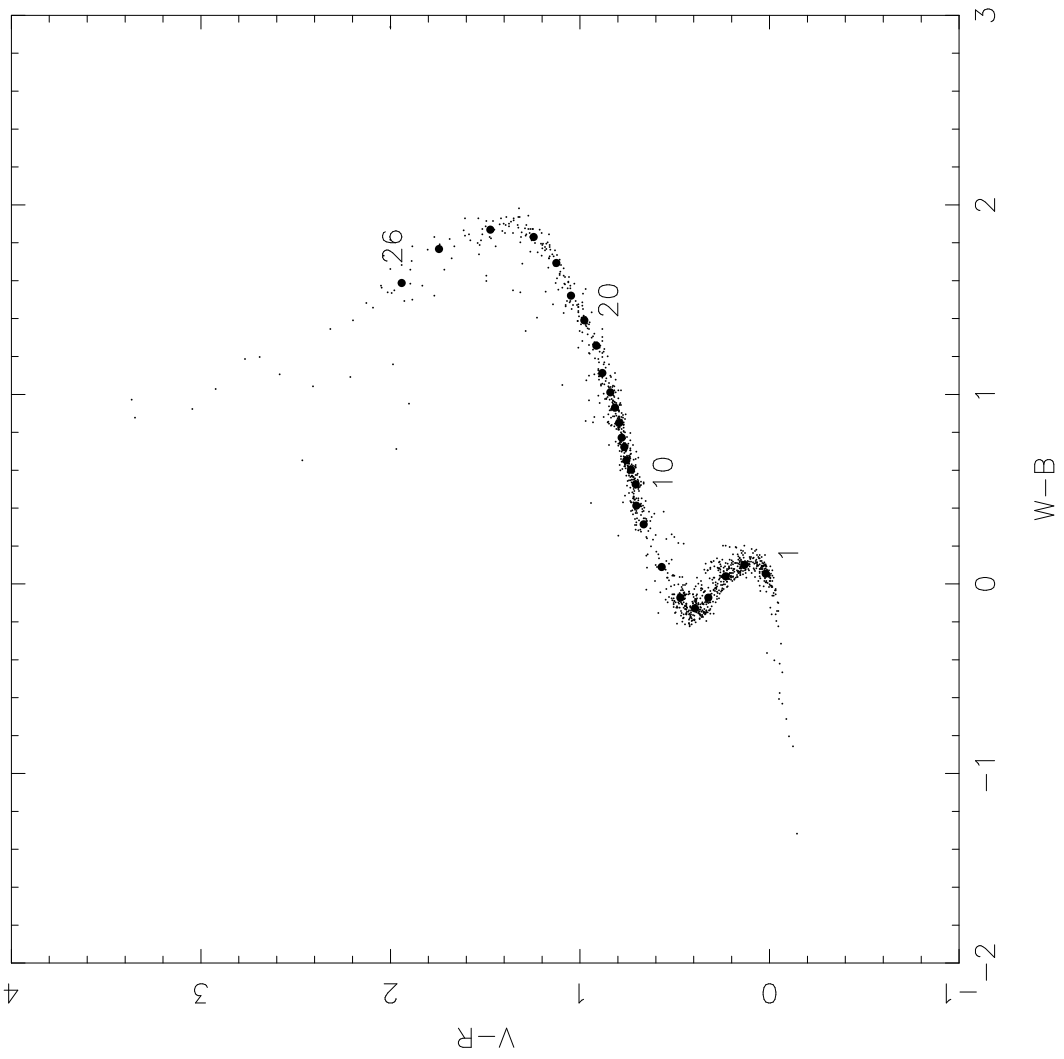

V−R
W−B
O
B
A
F
G
K
M
1
10
20
28
4
3
2
1
0
−1
−2
3
2
1
0
−1

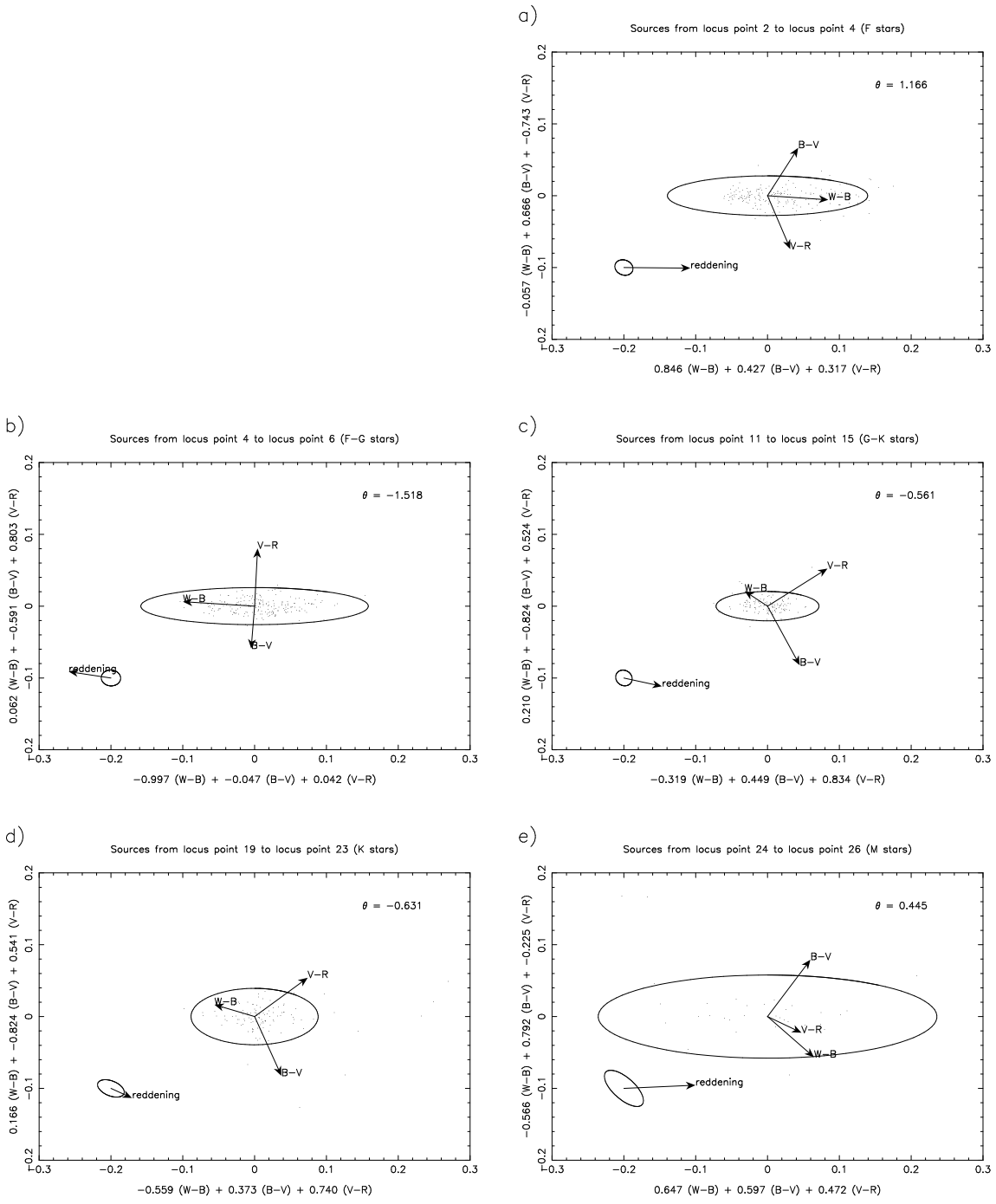

a)
Sources from locus point 1 to locus point 3 (B–A stars)
θ = −0.799
V–R
W–B
B–V
reddening
0.059 (W–B) + −0.697 (B–V) + 0.714 (V–R)
−0.181 (W–B) + 0.696 (B–V) + 0.694 (V–R)
b)
Sources from locus point 5 to locus point 7 (F stars)
θ = 1.101
B–V
W–B
V–R
reddening
−0.042 (W–B) + 0.652 (B–V) + −0.756 (V–R)
0.779 (W–B) + 0.494 (B–V) + 0.383 (V–R)
c)
Sources from locus point 7 to locus point 9 (F–G stars)
θ = −1.193
V–R
W–B
B–V
reddening
0.096 (W–B) + −0.594 (B–V) + 0.798 (V–R)
−0.907 (W–B) + 0.276 (B–V) + 0.315 (V–R)
d)
Sources from locus point 12 to locus point 16 (G–K stars)
θ = −0.652
V–R
W–B
B–V
reddening
0.202 (W–B) + −0.784 (B–V) + 0.586 (V–R)
−0.453 (W–B) + 0.455 (B–V) + 0.766 (V–R)
e)
Sources from locus point 20 to locus point 24 (K stars)
θ = −0.844
V–R
W–B
B–V
reddening
0.053 (W–B) + −0.721 (B–V) + 0.690 (V–R)
−0.573 (W–B) + 0.543 (B–V) + 0.612 (V–R)
f)
Sources from locus point 26 to locus point 28 (M stars)
θ = 0.395
B–V
V–R
W–B
reddening
−0.511 (W–B) + 0.826 (B–V) + −0.235 (V–R)
0.621 (W–B) + 0.545 (B–V) + 0.562 (V–R)

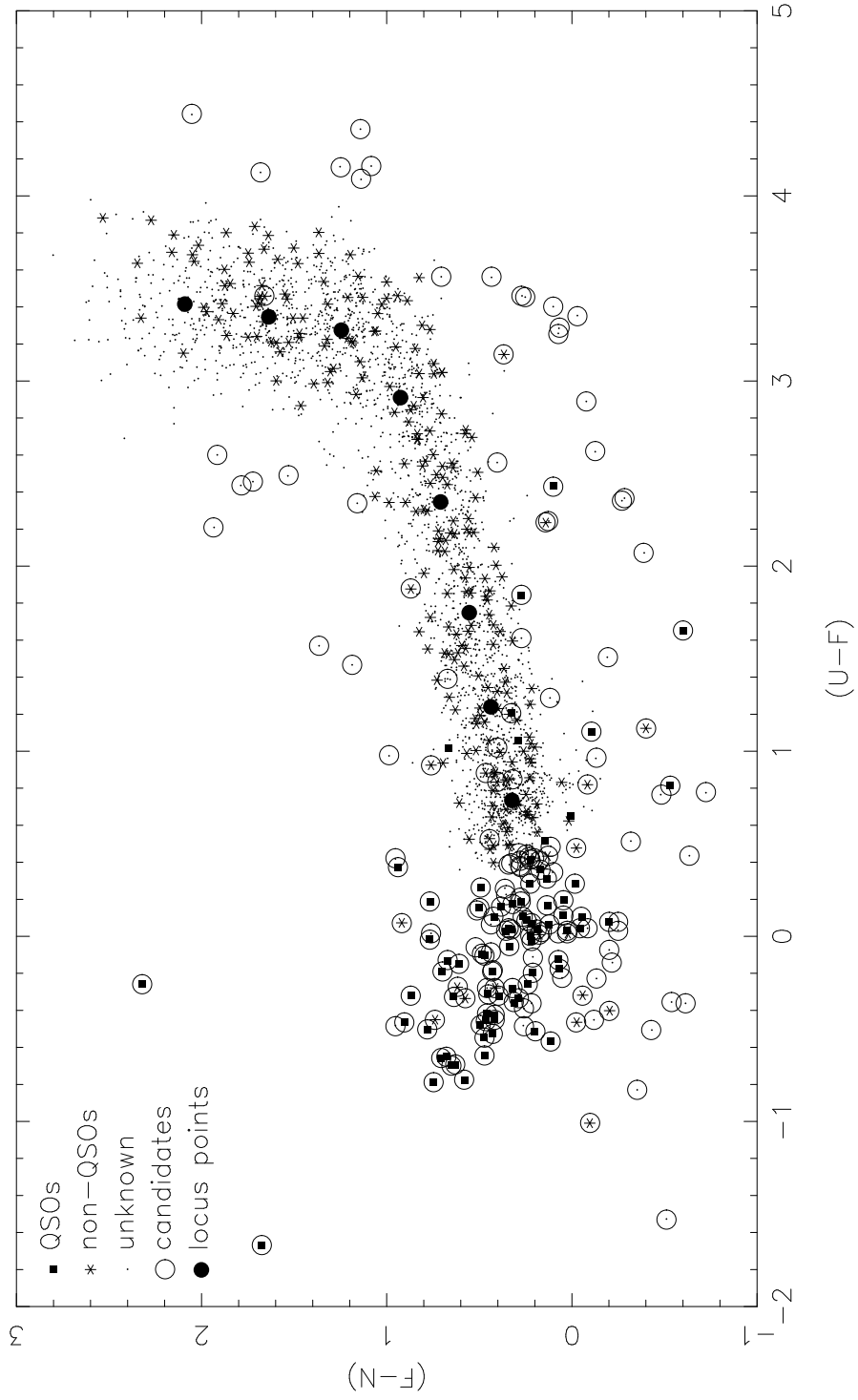

Number of Stars
0
100
200
300
400
500
0
2
4
6
8
Position Along Locus (k) in Magnitudes


Sources from locus point 12 to locus point 16 (G–K stars)
MINOR AXIS
σ = 0.010
Number of Stars
0
50
100
150
200
−0.03
−0.02
−0.01
0
0.01
0.02
0.03
0.202 (W–B) + −0.784 (B–V) + 0.586 (V–R)


MAJOR AXIS
σ = 0.028
Number of Stars
0
50
100
150
200
−0.1
−0.05
0
0.05
0.1
0.15
−0.453 (W–B) + 0.455 (B–V) + 0.766(V–R)


26
20
10
1
V−R
W−B
−1
0
1
2
3
4
−2
−1
0
1
2
3


a)
Sources from locus point 2 to locus point 4 (F stars)
θ = 1.166
B–V
W–B
V–R
reddening
−0.057 (W–B) + 0.666 (B–V) + −0.743 (V–R)
0.846 (W–B) + 0.427 (B–V) + 0.317 (V–R)
b)
Sources from locus point 4 to locus point 6 (F–G stars)
θ = −1.518
V–R
W–B
B–V
reddening
0.062 (W–B) + −0.591 (B–V) + 0.803 (V–R)
−0.997 (W–B) + −0.047 (B–V) + 0.042 (V–R)
c)
Sources from locus point 11 to locus point 15 (G–K stars)
θ = −0.561
V–R
W–B
B–V
reddening
0.210 (W–B) + −0.824 (B–V) + 0.524 (V–R)
−0.319 (W–B) + 0.449 (B–V) + 0.834 (V–R)
d)
Sources from locus point 19 to locus point 23 (K stars)
θ = −0.631
V–R
W–B
B–V
reddening
0.166 (W–B) + −0.824 (B–V) + 0.541 (V–R)
−0.559 (W–B) + 0.373 (B–V) + 0.740 (V–R)
e)
Sources from locus point 24 to locus point 26 (M stars)
θ = 0.445
B–V
V–R
W–B
reddening
−0.566 (W–B) + 0.792 (B–V) + −0.225 (V–R)
0.647 (W–B) + 0.597 (B–V) + 0.472 (V–R)


QSOs
non−QSOs
unknown
candidates
locus points
(F−N)
(U−F)
3
2
1
0
−1
−2
−1
0
1
2
3
4
5