

# The Algorithm to Select Outliers From the Stellar Locus in Three Colors

Heidi Jo Newberg

Fermi National Accelerator Laboratory, Box 500, Batavia, IL 60510

Electronic mail: heidi@fnal.gov

## 1. Overview

Given a parameterization of the locus of stars in SDSS colors, we would like to determine whether an individual catalog entry is consistent with being a star. Since there are so many more stars than QSOs in the region containing the stellar locus, and since the locus of QSOs is diffuse, one cannot recover much of the stellar locus region of multicolor space by asking whether the object is consistent with being a QSO. If it is consistent with being a star, it is probably a star.

For the individual catalog entry, we require the  $u' - g'$ ,  $g' - r'$ , and  $r' - i'$  colors. Additionally, we must know the variances and the covariances in each color. The target selection algorithm currently assumes that the errors in each filter are free from systematics, so that  $cov(a - b, b - c) = -var(b)$ . The variances for a given color is given by  $var(a - b) = var(a) + var(b)$ .

The parameterized locus contains a set of ordered locus points, going from the blue end to the red end. Each locus point is described by a  $(u' - g', g' - r', r' - i')$  position, along with an axial unit vector,  $\hat{k}$ , a position angle,  $\theta$ , and a major and minor axis,  $a_l, a_m$ , which describe an elliptical cylinder. Objects which lie within this cylinder, and which are between the planes which are equidistant from the current point and the two adjacent points, are considered to be within the stellar locus. In this way we build up a tube of elliptical cross section which can wind around in multicolor space.

In some cases, an object might not be measured in all filters, resulting in incomplete knowledge of the position in color space. For example, if an object was detected in  $g'$ , but was below the detection threshold in  $u'$ , then we know that  $u' - g' > u'_{lim} - g'$ . If the  $u'$  measurement was missed due to a defect in the image, then nothing would we known about the  $u' - g'$  color. In general, we either know the color and variances of an object; or we know a limiting color, and the variance is set to a flag which distinguishes between an upper limit, a lower limit, or no knowledge of that color. In this case, the algorithm tries to determine if there is any allowed value of the unknown colors which places the object within the stellar locus. If so, then the object is considered to be consistent with a galactic star and is not flagged as a candidate. We require that at least one color out of three be known.

The ends of the stellar locus are described by half ellipsoids which fit onto the ends of the tube. The center of the ellipsoids are constrained to lie on the axis of the stellar locus. A single input parameter determines position of each ellipsoid center along this axis ( $k_{blue}$  for the blue end and  $k_{red}$  for the red end). The axes of the ellipsoids are aligned with the axes of the elliptical cylinders

associated with the first (and last) locus points. Two of the axis widths are given by the major and minor axes of the closest locus point interior to the locus. The width along the locus,  $a_k$ , is a separate input parameter for each end. In the limit that  $a_k = 0$ , the locus ends with a plane rather than an ellipsoid.

The input locus parameterization describes the extent of the locus, not including any photometric errors in the measurements of stars. An input parameter,  $N_{\sigma_{err}}$ , determines how many sigma of the errors in an individual point should be used to “convolve” with the parameterized stellar locus. Before determining whether an object is consistent with being in the stellar locus, the widths of the locus and the ending ellipsoids are increased by adding the locus widths with the estimated errors in the axial directions in quadrature. This is a correct treatment if the density across the stellar locus and the error statistics are both Gaussian. Any limiting colors must have been previously adjusted to include any uncertainty in the knowledge of the limit, since variances along an axis with a limit will be assumed to be zero.

## 2. Finding the closest locus point

The first step is to determine which locus point is closest to the data point. We do this by looking at the Euclidean distance in color space, not taking into account any variation in the width of the locus, to simplify the computation. In the case that one or two of the colors are not known uniquely, the distance to each locus point is calculated after first moving the data point as close as is allowed by the color limits.

If two locus points are equally distant from the datapoint, the redder one is chosen. If the closest locus point is not between the centers of the ending ellipsoids, then the closest locus point is reassigned to be the first one interior to the ellipsoid centers.

Once the closest locus point is determined, we do not consider any other locus point. The convolution is done only on the cylinder associated with this one locus point, and the object is considered to be consistent with a star if the color is within that cylinder and interior to the ellipsoids at the ends of the locus.

## 3. Handling errors in individual objects

First, we estimate the error ellipse in the  $(l, m)$  plane. The variance, covariance matrix for this plane is given by:

$$V_{ab} = \sum_x \sum_y \frac{\partial a}{\partial x} \frac{\partial b}{\partial y} S_{xy},$$

where  $\mathbf{V}$  is the (two dimensional) covariance matrix in the  $l, m$  plane,  $\mathbf{S}$  is the (three dimensional) covariance matrix in color space, and the sums are over the three coordinates:  $u' - g', g' - r'$ , and

$r' - i'$ . The error ellipse is given by:

$$\vec{l}^T \mathbf{V}^{-1} \vec{l} = N_{\sigma_{err}}^2,$$

where  $\vec{l} = \hat{l} + m\hat{m}$ . Solving this, the ellipse parameters are:

$$a_{err}^2 = N_{\sigma_{err}}^2 \frac{V_{ll} + V_{mm} \pm \sqrt{(V_{ll} - V_{mm})^2 + 4V_{lm}^2}}{2},$$

$$\tan \theta_{err} = \begin{cases} 0 & V_{lm} = 0, V_{ll} > V_{mm} \\ \infty & V_{lm} = 0, V_{mm} > V_{ll} \\ \frac{-(V_{ll} - V_{mm}) + \sqrt{(V_{ll} - V_{mm})^2 + 4V_{lm}V_{mm}}}{2V_{lm}} & V_{lm} \neq 0 \end{cases}$$

If one convolves two bivariate Gaussians (with elliptical cross sections), one obtains a third bivariate Gaussian with elliptical cross section. In this case, the resulting ellipse is given by:

$$\alpha l^2 + 2\beta lm + \gamma m^2 = d,$$

where

$$d = a_{errmaj}^2 a_{errmin}^2 + (a_l^2 a_{errmaj}^2 + a_m^2 a_{errmin}^2) \sin^2 \theta_{err} + (a_{errmaj}^2 a_m^2 + a_l^2 a_{errmin}^2) \cos^2 \theta_{err} + a_l^2 a_m^2$$

$$\alpha = a_{errmin}^2 \cos^2 \theta_{err} + a_{errmaj}^2 \sin^2 \theta_{err} + a_m^2$$

$$\beta = -\sin \theta_{err} \cos \theta_{err} (a_{errmaj}^2 - a_{errmin}^2)$$

$$\gamma = a_{errmin}^2 \sin^2 \theta_{err} + a_{errmaj}^2 \cos^2 \theta_{err} + a_l^2$$

This resulting ellipse can be described by:

$$a_{totmaj}^2 = a + b,$$

$$a_{totmin}^2 = a - b,$$

$$\tan \theta_{tot} = \begin{cases} 0 & b = 0, \alpha = a \\ \infty & b - \alpha + a = 0, b \neq 0 \\ \sqrt{\frac{\alpha - a + b}{b - \alpha + a}} & b - \alpha + a \neq 0, \theta_{err} >= 0 \\ -\sqrt{\frac{\alpha - a + b}{b - \alpha + a}} & b - \alpha + a \neq 0, \theta_{err} < 0 \end{cases},$$

where

$$a = \frac{(\alpha + \gamma)}{2} = \frac{a_{errmaj}^2 + a_{errmin}^2 + a_l^2 + a_m^2}{2},$$

$$b = \frac{\sqrt{(\alpha + \gamma)^2 - 4(\alpha\gamma - \beta^2)}}{2} = \frac{\sqrt{(a_{errmaj}^2 - a_{errmin}^2)^2 + 2(a_{errmaj}^2 - a_{errmin}^2)(a_l^2 - a_m^2)(\cos^2 \theta_{err} - \sin^2 \theta_{err}) + (a_l^2 - a_m^2)^2}}{2}.$$

Here,  $a$  and  $b$  are positive by definition. We need not worry about taking the square root of a negative number when computing the angle, since  $a - b \leq \alpha \leq a + b$  for all cases. If  $b = 0$ , then the resulting ellipse is circular, so we might as well make  $\theta = 0$ .

We now can redefine the  $\hat{l}, \hat{m}$  axes associated with the closest locus point to describe this new elliptical cross section. The value of  $\theta_{tot}$  must be added to the original angle describing the locus ellipse, since we have done this with respect to the  $\hat{l}$  axis. When adding, we must be sure to keep the new value of  $\theta$  in the allowed range:  $-\frac{\pi}{2} < \theta \leq \frac{\pi}{2}$ .

In addition to increasing the width of the excluded locus, we increase  $a_k$  on the ends of the locus in a similar manner:

$$\begin{aligned} var(k) &= \sum_x \sum_y \frac{\partial k}{\partial x} \frac{\partial k}{\partial y} S_{xy}, \\ a_{ktot} &= \sqrt{a_k^2 + N_{\sigma_{err}}^2 var(k)}, \end{aligned}$$

where the  $k$  in this equation is not for the locus point which is closest to the datapoint, but instead for the locus point which is closest to the respective endpoints (subject to being between the two endpoints).

In the remainder of this paper we will use  $a_l, a_m, a_k$ , and  $\theta$  to refer to the derived  $a_{totmaj}, a_{totmin}, a_{ktot}$  and  $\theta \pm \theta_{tot}$ . Likewise, the  $\hat{l}, \hat{m}$  unit vectors are in the new coordinate system.

#### 4. Dealing with non-detections in some filters

If not all of the colors of the datapoint are known, then we are free to find the allowed point in color space that is most likely to be in the locus. This is *not* in general the same as finding the colors that are closest to the locus point in the Euclidean sense, since the parameterized region is an elliptical cylinder, not a sphere centered at the locus point. We instead wish to minimize the  $r^*$  distance to the locus, where  $r^*$  is given by:

$$r^* = \sqrt{\left(\frac{l}{a_l}\right)^2 + \left(\frac{m}{a_m}\right)^2},$$

where  $l \equiv \Delta \vec{r} \cdot \hat{l}$ ,  $m \equiv \Delta \vec{r} \cdot \hat{m}$ ,  $\Delta \vec{r} \equiv \vec{r} - \vec{r}_p$ , and  $\vec{r}_p$  is the position in color space of the closest point.

First, let's solve the case where only the  $u' - g'$  color, is unknown. To make the equations easier to read, I will use the notation  $\Delta \vec{r} \equiv (\Delta x, \Delta y, \Delta z)$ ,  $l = l_x \hat{x} + l_y \hat{y} + l_z \hat{z}$ ,  $m = m_x \hat{x} + m_y \hat{y} + m_z \hat{z}$ . The solution is:

$$\Delta x = \begin{cases} -\frac{\{(a_m^2 l_x l_y + a_l^2 m_x m_y) \Delta y + (a_m^2 l_x l_z + a_l^2 m_x m_z) \Delta z\}}{a_m^2 l_x^2 + a_l^2 m_x^2} & a_m^2 l_x^2 + a_l^2 m_x^2 \neq 0 \\ -\frac{l_y \Delta y + l_z \Delta z}{l_x} & a_m^2 l_x^2 + a_l^2 m_x^2 = 0, l_x \neq 0 \\ -\frac{m_y \Delta y + m_z \Delta z}{m_x} & a_m^2 l_x^2 + a_l^2 m_x^2 = 0, m_x \neq 0 \\ 0 & a_m^2 l_x^2 + a_l^2 m_x^2 = 0, l_x = 0, m_x = 0 \end{cases}.$$

The first case above is the result of the minimization process. If the denominator is zero, but  $l_x \neq 0$ , then it follows that  $a_m = 0$ . Also, either  $a_l = 0$  or  $m_x = 0$ . In the first case, we will only be able to find a color point in the locus if we can fortuitously set  $m = l = 0$  by moving along the  $x$  axis. In the second case, we cannot affect the magnitude of  $m$ . Either case is optimized by setting  $l = 0$ . If the denominator is zero, but instead  $l_x = 0$ , then we still need either  $a_l = 0$  or  $m_x = 0$ . The first case we cannot affect the distance along the major axis, so we set the distance along the minor axis to zero. The second case gives us  $\hat{x} = \hat{k}$ , so we might as well set  $\Delta x = 0$ .

If the value of  $x$  is completely unknown, then we use this calculated  $\Delta x$  to determine whether the datapoint is consistent with being a star. If the value of the color  $x$  is a limit, then we must ask whether the computed value is within the limits. If it is not, then we instead use the limiting  $x$  value to determine whether the datapoint is consistent with being a star.

Next, we tackle the case where both  $u' - g'$  and  $g' - r'$  are unknown. When two of the colors are unknown, one can find values that will put the point exactly on the line  $\vec{r} - \vec{r}_p = \lambda \hat{k}$ , where  $\lambda$  is a free parameter. We can solve for  $\lambda$ , and then the two unknown colors using:

$$\begin{aligned} \lambda &= \begin{cases} \frac{\Delta z}{k_z}, & k_z \neq 0 \\ 0, & k_z = 0, \end{cases} \\ \Delta y &= \lambda k_y, \\ \Delta x &= \lambda k_x. \end{aligned}$$

If  $k_z = 0$ , then the line down the center of the locus is in the  $x, y$  plane, so there are many values of  $x$  and  $y$  which will be on it. We arbitrarily choose the one on the locus point, since we know that this value is also within the  $k$  limits of the stellar locus.

The existence of limiting values in  $x$  and  $y$  make this a little trickier. First, we calculate the value of  $\Delta y$  which places the point exactly on the center of the locus. In the case that this value is not consistent with the  $y$  limits, we assign  $\Delta y = y_{lim} - y_p$ . Since we now have only one missing color, we can use the procedure outlined above (the case where only the  $u' - g'$  color is unknown) to calculate the optimal value of  $\Delta x$ . If a limit was not reached, one can verify that the computed  $\Delta x$  will be the same as if we had used  $\Delta x = \lambda k_x$ . If this computed value of  $\Delta x$  within the allowed limits, then we have done the best we can. If it is not, then we set  $\Delta x = x_{lim} - x_p$  and then compute the optimal value of  $\Delta y$  given  $\Delta x$  and  $\Delta z$ . The optimal value of  $\Delta y$ , given  $\Delta x$  and  $\Delta z$ , can be computed by reassigning the axes ( $x \rightarrow y, y \rightarrow z, z \rightarrow x$ ) in the equation that optimizes  $\Delta x$  given  $\Delta y$  and  $\Delta z$ .

The above equations and their permutations ( $x, y, z \rightarrow y, z, x$  and  $x, y, z \rightarrow z, x, y$ ) are used to determine the values of  $\Delta x$ ,  $\Delta y$ , and  $\Delta z$  which are most likely to be included in the stellar locus, given the input limits on these quantities. Once we have this coordinate in three dimensions, we

can ask whether it is within the parameterized locus. We must have  $r^* \leq 1$ , where

$$r^* = \begin{cases} \sqrt{\left(\frac{l}{a_l}\right)^2 + \left(\frac{m}{a_m}\right)^2} & a_l > 0, a_m > 0 \\ \frac{l}{a_l} & a_l > 0, a_m = 0, m = 0 \\ 0 & l = m = 0 \\ \infty & a_l = 0, l \neq 0 \\ \infty & a_m = 0, m \neq 0 \end{cases}$$

The point must also be within the ellipsoids at the ends of the locus. If  $\Delta k < 0$ , then we already know we are within the red end limit, since the locus point is guaranteed to be within the locus. Similarly, if  $\Delta k > 0$ , we are guaranteed to be within the blue end limit. So, we need only check one of the ellipsoids. We ignore the fact that the locus may curve around in  $x, y, z$  coordinates, and only look at the  $l, m, k$  coordinates. We are within the ends of the locus if  $k_{blue\_limit}(l, m) < k < k_{red\_limit}(l, m)$ , where:

$$k_{blue\_limit} = k_{blue} - a_{k\ blue} \sqrt{1 - r_{blue}^*{}^2}$$

$$k_{red\_limit} = k_{red} - a_{k\ red} \sqrt{1 - r_{red}^*{}^2}.$$

Here, the values of  $r^*$  are computed using  $a_l$  and  $a_m$  for the end ellipses, but values of  $l$  and  $m$  calculated from the closest locus point.

## 5. End conditions

If the previously calculated optimal values of  $\Delta x, \Delta y$ , and  $\Delta z$  put us within the locus, then we are done. If we could not find values for the coordinates that made  $r^* \leq 1$ , then we are done. However, if the values failed only the end conditions, then there is a chance we could still put the datapoint in the locus by minimizing the distance to the center of the end ellipsoid:

$$r^{**} = \sqrt{\left(\frac{l'}{a'_l}\right)^2 + \left(\frac{m'}{a'_m}\right)^2 + \left(\frac{k' - k'_{end}}{a_k}\right)^2}$$

rather than the center of the cylinder associated with the closest locus point. Here,  $k_{end}$  is either  $k_{blue}$  or  $k_{red}$ , depending on whether  $\Delta k < 0$  or  $\Delta k > 0$ . The primes indicate that the values are measured with respect to the position in color space of the locus point which is closest to the center of the end ellipsoid. The  $a_l$  and  $a_m$  widths are for the end ellipsoid, not the cylinder associated with the closest locus point.

We now calculate for the datapoint a new optimal color which minimizes  $r^{**}$ . If only  $x$  is unknown, we compute  $\Delta z'$  and  $\Delta y'$  using the definition  $\Delta \vec{r}' \equiv \vec{r} - \vec{r}_e \equiv (\Delta x', \Delta y', \Delta z')$ , and  $\vec{r}_e$  is

the position in color space of the locus point closest to the center of the end ellipsoid (subject to  $k_{blue} < k_e < k_{red}$ ). The value of  $\Delta x'$  which minimizes  $r^{**}$  is

$$\Delta x' = \begin{cases} -\frac{a'^2_m a'^2_k (l'_y \Delta y' + l'_z \Delta z') l'_x + a'^2_l a'^2_k (m'_y \Delta y' + m'_z \Delta z') m'_x + a'^2_l a'^2_m (k'_y \Delta y' + k'_z \Delta z' - k'_{end}) k'_x}{a'^2_m a'^2_k l'^2_x + a'^2_l a'^2_k m'^2_x + a'^2_l a'^2_m k'^2_x} & D_1 \neq 0 \\ -\frac{a'^2_k (l'_y \Delta y' + l'_z \Delta z') l'_x + a'^2_l (k'_y \Delta y' + k'_z \Delta z' - k'_{end}) k'_x}{a'^2_k l'^2_x + a'^2_l k'^2_x} & D_1 = 0, a'^2_m = 0, m'_x = 0, a'^2_l \neq 0 \\ -\frac{m'_y \Delta y' + m'_z \Delta z'}{m'_x} & D_1 = 0, a'^2_m = 0, m'_x \neq 0, a'^2_l \neq 0 \\ -\frac{a'^2_m (l'_y \Delta y' + l'_z \Delta z') l'_x + a'^2_l (m'_y \Delta y' + m'_z \Delta z') m'_x}{a'^2_m l'^2_x + a'^2_l m'^2_x} & D_1 = 0, a'^2_m > 0 \\ \frac{k'_x (k'_y \Delta y' + k'_z \Delta z')}{1 - k'^2_x} & D_1 = 0, a'^2_l = 0, m'_x = 0, k'_x \neq 1 \\ 0 & D_1 = 0, l'_x = 1 \end{cases}$$

where  $D_1 \equiv a'^2_m a'^2_k l'^2_x + a'^2_l a'^2_k m'^2_x + a'^2_l a'^2_m k'^2_x$ . The first case is the formal result of the minimization. The second case is for  $a'^2_m = 0$  and  $m'_x = 0$ ; in this case we cannot affect the  $m$  value, so we might as well find the best position within the  $l, k$  ellipse. If instead we have  $m'_x \neq 0$ , we instead put the datapoint on the  $m = 0$  plane. If  $D_1 = 0$  but  $a'^2_m > 0$ , then we must have  $a'^2_k = kx = 0$ , since  $a'^2_l > a'^2_m$ . If  $a'^2_l = 0$ , then the best we can do is move the point onto the  $k$  axis, assuming  $k'_x \neq 1$ , so that we can move perpendicular to the axis. All cases where  $D_1 = 0$  and  $l'_x = 1$  reduce to  $\Delta x' = 0$ , and this sweeps up all of the cases not covered by the other criteria. Again, if this computed value for  $\Delta x'$  violates a limit in  $x$ , then we reassign  $x$  to the limit.

If both  $x$  and  $y$  are unknown, then we start by calculating the values which minimize  $r^{**}$ . We deal with limits in the identical way as we did when we were minimizing  $r^*$ . That is, we figure out the optimal value of  $x$  and  $y$  together, but only assign the  $y$  value. Then we figure out the optimal value of  $x$  given that  $y$  value and the fixed value of  $z$ . This way we can deal with the limits in a sensible way. The optimal value of  $y$  is given by:

$$\Delta y' = \begin{cases} \frac{(k'_y k'_z a'^2_k + m'_y m'_z a'^2_m + l'_y l'_z a'^2_l) \Delta z' - (l'_x m'_z a'^2_m - m'_x l'_z a'^2_l) k'_{end}}{a'^2_k k'^2_z + a'^2_m m'^2_z + a'^2_l l'^2_z} & D_2 \equiv a'^2_k k'^2_z + a'^2_m m'^2_z + a'^2_l l'^2_z \neq 0 \\ k'_y k'_{end} & D_2 = 0 \end{cases}$$

The first option is the formal minimization of  $r^{**}$ . If  $k'_z = 0$ , then the center of the ellipsoid is in the  $x, y$  plane. Therefore, we try to place the datapoint on that center (or as close as we can get, given that  $\Delta z'$  might not be zero). If the denominator is zero, then  $a'^2_m = 0$ . This is true because we cannot have  $k'_z = 1, l'_z = 0, m'_z = 0$  due to the definition of our system. So, either  $a'^2_l$  or  $a'^2_m$  must be zero. Either way,  $a'^2_m = 0$ . If  $k'_z \neq 0$ , we must have  $a'^2_k = 0$  as well. Additionally, either  $a'^2_l = 0$  (the endpoint is the only hope) or  $l'_z = 0$  ( $\hat{l}$  is in the  $x, y$  plane, so the endpoint is still the best choice, if we can get there).

As before, we then adjust  $\Delta y'$  if it is outside the allowed limits, then use our equations for only  $\Delta x'$  missing to find the optimal  $x'$ . If the computed value of  $x'$  is outside the limits, then we set  $x'$  to the limit, and recompute the optimal value of  $y'$ .

I cannot think of a case in which the new position in the interior side of the endplane at  $k_{end}$ . So, all I have to do is figure out if the new position satisfies  $r'^* \leq 1$  and is interior to the  $k$

limits on the outer surface of the ellipsoid. If it is, then it is consistent with being a star, otherwise it is not.