Analysis of Molecular Clouds Using the Spectral Correlation Function

Erik W. Rosolowsky

Harvard Center for Astrophysics, Cambridge, MA 02181

Swarthmore College, Swarthmore, PA 19081

Received ______________________; accepted ______________________

ABSTRACT

The Spectral Correlation Function (SCF) is an algorithm that calculates the similarity between spectra in maps of molecular clouds. The function was developed in order to provide another diagnostic in examining the maps of clouds. When the size of these maps is larger than a few hundred pixels, it becomes impossible to analyze them rigorously by human observation. Thus, the function is intended to provide more quantitative data in a fashion that is easily interpreted.

The primary goal of the project was to measure the correlation between spectra in different data sets and to evaluate how well Magneto-Hydrodynamic simulations of molecular clouds model clouds in the interstellar medium. In the absence of any such data cubes, other tests were devised to analyze the results. These tests indicate that the SCF algorithm can distinguish between data sets with good correlation between their spectra and those with poor correlation.

1. Background

When you look up at night, some 6000 stars in the galaxy greet the naked eye with varying brightnesses and colors. The particularly keen-eyed observer can even notice some nebulosity associated with such features as Orion's sword. These nebulae are gas and dust that reflect the ambient starlight thereby allowing an observer to see their massive bulk. These illuminated features represent a small fraction of the vast amounts of gas and dust in the galaxy that is not contained in stars. Apart from extinction caused by dust or the bright Hα radiation of an HII region, very little of the interstellar medium can be seen in the optical wavelengths.

Much of the interstellar medium (ISM) is cool with temperatures from 10 K up to 70 K (Dyson and Williams, 1980). These low temperatures indicate that low energy interactions will dominate in the material and thus low energy photons will be produced ($kT = h\nu \Rightarrow \nu \sim 10^{11}\ Hz$). The material, therefore, is best viewed in the radio band. The majority of the transitions at these low energies are the rotational and, more rarely, vibrational transitions of molecules. Although hydrogen in the form of H_2 dominates the ISM in numbers, the molecule has no dipole moment of charge or mass; thus the rotational transitions are forbidden. To excite vibrational states of H_2 significantly, the temperatures required are in excess of those found in the typical cool ISM. Consequently, radio observations of the cool ISM focus on the transitions of significantly less populous molecules such as CO. When the sky is searched for photons with these characteristic frequencies, enormous structures spanning tens of degrees are discovered (Meyers, 1987). These structures are found to have temperatures on order of 10 to 20 K and have number densities of $10^3\ cm^{-3}$ (Dyson and Williams, 1980), a factor of 10^3 above the average density. In order to determine the structure and kinematics of the clouds, more data must be considered.

Two spatial dimensions of the clouds can be measured by the pointing of the radio telescope. In order to supplement these, the third dimension of velocity is added by taking spectra in a grid pattern on the sky. By examining the spectra, the kinematics of the cloud can be deduced. A good example of this procedure is found in Pound and Goodman's analysis of the Ursa Major cloud complex (1997). In small sets of data, with only a few pixels on each side, the behavior of the gas can be deduced by human observation. When the data sets become large, with hundreds of pixels on a side, every spectrum cannot be examined. Numerical methods of analysis must be developed to automate the process and aid the human observer. Most of the methods to date center around reducing the data contained in the velocity dimension to a single value allowing for two-dimensional maps to be considered. This reduction constitutes an immense loss of information; thus new methods should be developed so as to minimize the losses.

2. Introduction

The existing methods of analyzing maps of spectra have been applied to molecular clouds which have been found to exhibit specific observational properties. For example, Larson (1981) discovered a relationship between internal velocity dispersion (Δv) and the size of the cloud (R), following a power law $\Delta v \propto R^a$ with $a = 0.38 \pm 0.14$. Subsequent studies have confirmed and refined this relationship (Fuller & Myers, 1992 and references therein) and separated the measurement of the velocity dispersions into thermal and non-thermal components. The non-thermal component is the only component to be affected by said power law relationship; hence the increase is often explained as turbulence with velocity dispersion increasing as a power law of the size.

Recent computer simulations have generated models which can reproduce the quantitative relationships developed in the observation of molecular clouds. This modeling has been accomplished with good success For example, Gammie and Ostriker (1996), Passot *et. al* (1995), and Vázquez-Semadeni (1996) all generate models which exhibit these relationships. However, when the maps produced by these models are observed, it is not clear that structures similar to those found in real clouds are formed (for example, Falgarone (1994)). The structure generated by these models might be another solution that yields the same average properties over the map as are found in real observations. This mimicking can be done because the relations do not uniquely define a cloud topology.

The purpose of the SCF is to invent a process like Larson's that quantifies a velocity structure, thereby allowing for another basis of comparison between model data and real data. The SCF is intended to be used as a discriminant, where data that are indistinguishable from real data in the properties measured by the SCF are those that best model the real ISM.

3. Procedure

The primary focus of the project was to develop a function that represents the correlation between two spectra and generate a computer algorithm that efficiently calculates the correlation. This task required several parts, which are described below. The first part was to calculate the correlation efficiently, so that large maps of spectra could be analyzed in a reasonable amount of time. Next, the algorithm was applied to simple, artificially generated data cubes so that the output could be interpreted. Finally, the algorithm was used to analyze data cubes from observations.

The basis for this plan can be found in Goodman (1997).

3.1. Efficient Optimization

The correlation function, around which the entire algorithm centers, refers to the deviation between two spectra, $T^*_{A,1}(v)$ and $T^*_{A,0}(v)$. As implied by the notation, these spectra are in the units of antenna temperature versus velocity. The deviation function, δ, is also a function of velocity and is given by:

$$[\delta_{1,0}(v)]^2 = [sT^*_{A,1}(v-\tau) - T^*_{A,0}(v)]^2 \tag{1}$$

This equation contains the two free parameters s, a scaling factor, and τ, a lag, which are to be adjusted so that the deviation function is minimized.

The problem of determining how well two spectra can be correlated is tantamount to minimizing this deviation. Thus, an appropriate metric for determining the minimization of $[\delta(v)]^2$ is required. In this case, the chosen metric is the integral over velocity of the deviation, i.e. $D(T_1, T_0) \equiv \int [\delta_{1,0}(v)]^2$. The minimization of this integral by adjusting the parameters s and τ will yield the most closely correlated spectra. It is these two parameters which are interesting, for their values, when the deviation is minimized, will indicate the fashion in which the spectra are correlated. For example, if the value of τ is close to 0 for a certain correlation, the two spectra should have similar velocity distributions. If the value of s is close to 1 for the correlation, the two spectra should have similar temperatures over the velocity range.

In minimizing maps made from a large number of spectra, it is important that the minimization of these parameters be done efficiently and accurately. In order to understand what simplifying assumptions can be made, the deviation function must be studied for insight. The immediate difficulty lies with τ because it is a parameter in the argument of T^*_A. The fact that T^*_A is not defined analytically implies the following:

1. The value of $[\delta(v)]^2$ can only be evaluated for values of τ such that $T^*_A(v-\tau)$ is defined by the original spectrum. In other words if $v' = v - \tau$, then v' must be a velocity for which an antenna temperature is measured in the original spectrum.

2. Several minimization routines require the value of $\partial[\delta(v)]^2/\partial\tau$. Any accurate partial derivative of $[\delta(v)]^2$ with respect to τ requires evaluating $T_A^*(v-\tau)$ at several points. For noisy functions, the number of points evaluated must be quite high in order for the partial to be of any use to a numerical optimization routine.

3. The function $T_A^*(v)$ has many local extrema and thus any minimization routines, in seeking a minimum of τ, will find a local minimum. In order to insure this is the absolute minimum, most routines check to see if a randomly selected initial point will cause the algorithm to return to the already located minimum. If not, the values of the function at the two points are compared and the lower one is selected for the same displacement treatment. By numerous iterations of the program, this checking procedure will eventually discover the minimum. However, for typical spectra, this would require a large number of iterations, far more than most programs are intended to use.

Fortunately, it is much easier to find the appropriate values of s. Viewed as a function of s, the deviation can be expanded algebraically:

$$D(T_1,T_0) \equiv \int[\delta(v)]^2 dv = s^2\int[T_{A,1}^*(v-\tau)]^2 dv - 2s\int T_{A,1}^*(v-\tau)T_{A,0}^*(v)dv + \int[T_{A,0}^*(v)]^2 dv \qquad (2)$$

When seen this way, it is relatively easy to note that for each possible value of τ there exists only one value of s such that the function $D(T_1,T_0)$ has a local minimum. This value is given by equating the derivative with respect to s with zero:

$$s_{min} = \frac{\int T_{A,1}^*(v-\tau)T_{A,0}^*(v)dv}{\int[T_{A,1}^*(v-\tau)]^2 dv} \qquad (3)$$

Because there is only one local minimum, s_{min}, it is possible to minimize the function $D(T_1,T_0)$ with respect to s first and then with respect to τ. Reversing the order of this minimization is not necessarily valid for all possible $T_A^*(v)$. Based on the above considerations, the most efficient algorithm developed was to calculate s_{min} for all reasonable values of τ and then select the smallest value of $D(T_1,T_0)$ using these parameters.

3.2. Numerical Integration

The evaluation of the deviation, $D(T_1, T_0)$, requires the computation of the three integrals in equation (2). In addition, minimizing s requires the computation of the same integrals. There are many numerical integration routines available on different computer platforms. However, efficiency was a priority in designing the algorithm and the simplest numerical method was used. The spectral data are sets of paired numbers $(v_i, T^*_{A,i})$ which can be plotted as $T^*_A(v_i)$ to yield the usual spectrum. Most spectral data are in the special case where the velocity abscissae are evenly spaced: $v_{i+1} - v_i = \ \mathit{constant} \ \equiv dv$ for all i. In order to integrate an algebraic function of spectra over the velocity, the values of the function are summed and the resulting total is multiplied by dv.

$$\int f(v)dv \simeq \left(\sum_i f(v_i)\right) \cdot dv \tag{4}$$

This approximation is excellent, so long as the sampling of the data, dv, is smaller than the scale over which the function $f(v)$ varies appreciably. For sample data, like Gaussians, an integration accurate to 1 part in 10^7 requires but 2 samples per half width. Unfortunately, most spectral data is noisy and thus not slowly varying on a scale larger than the sampling. The integral under the curve, therefore, will be significantly different from the total derived using the summation approximation. Each of the spectra involved in the function $f(v)$ has an associated value σ representing the noise in the spectrum. This value is the root-mean-squared value of the noise of the signal where it should theoretically be zero and can be used as the error in the spectrum. Thus, the error in the integration routine can be approximated by calculating the errors in quadrature. In all cases, the integrals are of the form $g = \int f(v)dv = \int T_i(v)T_j(v)dv$. Thus, the inherent error in the calculation is:

$$\delta f(v) = \sqrt{T_j^2(v)\sigma_i^2 + T_i^2(v)\sigma_j^2} \tag{5}$$

$$\delta g = \int \delta f(v)dv \simeq \left(\sum_i \delta f(v_i)\right) \cdot dv \tag{6}$$

In the above equation, a “δ” preceding a function indicates that it is the error in that function. By means of example, the error incurred by noise in the spectra is about 1 part in 4 whereas the error incurred by the summation integration is less than 1 part in 25 for a spectrum with signal to noise value of 3.18 and a channel width of 0.05 km/s. Thus, the error due to the approximation can be neglected in light of the error due to noise in the spectrum.

3.3. Normalization

In order to aid in interpreting the results of the deviation calculation, the value of said calculations must be normalized. The function $D(T_1, T_0)$ was normalized to the interval [0,1] with 1 indicating perfect correlation and 0 meaning the minimum correlation in the following fashion:

$$S(T_1, T_0) = 1 - \sqrt{\frac{D(T_1, T_0)}{s^2 \int T_1^2(v)dv + \int T_0^2(v)dv}} \tag{7}$$

The normalization value in the denominator is chosen because it represents the maximum value of the deviation in the absence of absorption. The selection of this value can be understood by examining the second integral term in Equation 2. In the equation, the other two terms are positive and of a definite value for any given spectrum. The second term will be negative unless either of the spectra have negative values of $T_A^*(v)$. The spectra will be minimally correlated when the product of the two spectra is minimized over the entire range. The corresponds to having the peaks widely separated so that each peak is multiplied by the near-zero baseline of the other spectrum. In the limit of bad correlation, the second term will be zero, indicating the maximum deviation to be that found in the denominator of the radical in equation 7.

As mentioned before, the second term will be positive if there is a large contribution due to absorption. In this case, the final normalized deviation will be negative. This is because $\int [\delta(v)]^2 dv > s^2 \int T_1^2(v)dv + \int T_0^2(v)dv$. Negative results in the calculation of spectral correlation $S(T_1, T_0)$ can be understood as a result of this or, in regions of low signal to noise, as pathologies associated with the antenna temperature dropping below zero.

3.4. Applying the SCF

The above sections have described how to quantify the correlation between two individual spectra. The SCF, however, is intended to be applied to data cubes. A data cube is a two dimensional map of the sky with a spectrum at observed position. While these cubes can be thought of as maps in $\{x, y, v\}$ space, it is more helpful to think of them as having a spectrum at every point in a two dimensional map (See Figure 1).

The following procedure is used to generate maps of spectral correlation, lag, scaling and other parameters, given an input data cube.

1. Consider a data cube with size $X \times Y$ and spectra at each (x_i, y_j) position in the cube. For purposes

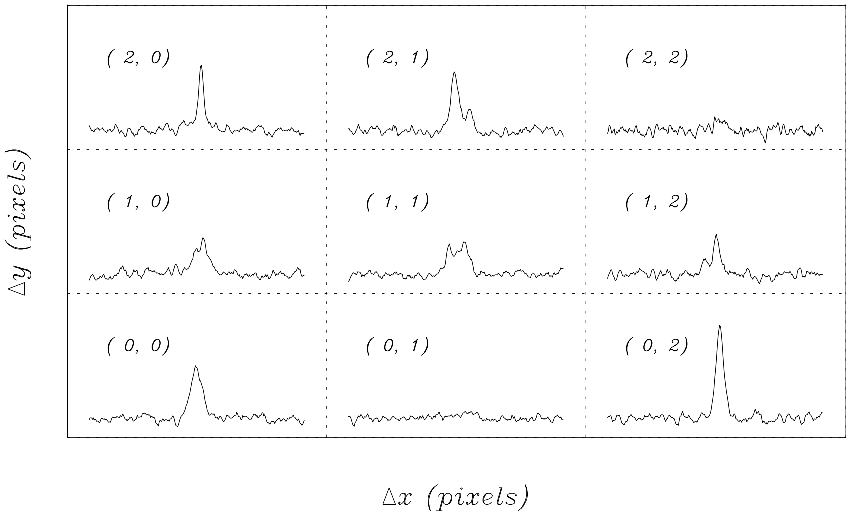

Fig. 1.— Sample section of a Data Cube with Pixel Coordinates

of the calculation, it is assumed that the spectra are taken on an evenly spaced grid. In the discussion to follow, (x_i, y_j) represent the real right ascension and declination coordinates of the cube and the coordinates (i, j) are the pixel coordinates in the data cube.

2. For each of these spectra, a Gaussian is fitted to each spectrum, allowing for the estimation of the line parameters. The parameters derived from the fit are peak antenna temperature (T_A^*), the half width of the Gaussian in velocity space (Δv), the center of the Gaussian (v_{LSR}), the integrated area under the Gaussian (T_{int}), generated with the summation integration (see above), and the rms noise of the spectrum (σ_{rms}). Spectra with a signal-to-noise ratio below a selected threshold are rejected and all correlation calculations are performed without these rejected data.

3. A test spectrum at (x_i, y_j) is selected to measure its correlation with nearby spectra.

4. The SCF requires two input parameters denoted by the variables r and q. r is the resolution of the SCF and represents the number of spectra for which the correlation with the test spectrum is measured. A resolution of r means to consider an $r \times r$ box centered on the test spectrum and to compare the test spectrum with every spectrum in the box. If desired, the results of the correlation calculations can be weighted by the distance away from the test spectrum.

   The parameter q is a measure of what velocity range is compared with the other spectra in the $r \times r$

box. All calculations discussed in the following section are to be performed only on the velocity range

$$v_{LSR}(i,j) - q\Delta v(i,j) \leq v \leq v_{LSR}(i,j) + q\Delta v(i,j) \tag{8}$$

The Gaussian fit parameters in Equation (8) are taken from the fit to the test spectrum at (x_i, y_j). All the integrations performed in the calculation of the deviation function and the SCF itself are taken over this velocity range as opposed to the full range of velocities within the spectral bandpass.

5. A comparison spectrum at (x_a, y_b) is selected with the provisions that $a \in [i-(r-1)/2, i+(r-1)/2]$, $b \in [j-(r-1)/2, j+(r-1)/2]$, $a \neq i$, and $b \neq j$.

6. Letting the spectrum at (x_a, y_b) equal $T_1(v)$ and the spectrum at (x_i, y_j) equal $T_0(v)$, the deviation function, $D(T_1, T_0)$, is minimized by adjusting s and τ. The deviation function is then normalized according to prescription in Equation 7.

7. The above step is repeated for each spectrum in the $r \times r$ box. The resulting values for $S(T_{a,b}, T_{i,j}), s_{a,b,i,j}$, and $\tau_{a,b,i,j}$ are averaged over all non-rejected values of a and b in the box. The resulting mean is then to the position at (x_i, y_j) as $S_{ij}, s_{ij}, \tau_{ij}$.

8. The deviation function is then minimized between the test spectrum and all comparison spectra holding the value of s at 1. This represents the correlation between the areas of the spectra because differences in area can not be compensated by the scaling parameter s. The averaged values of the SCF and the lag, τ, over all applicable a and b are referred to as S^l_{ij} and τ^l_{ij} respectively.

9. The deviation function is then minimized holding $\tau = 0$, in an attempt to explore velocity offsets alone. The values of the SCF and $s_{a,b,i,j}$ in this case are referred to as S^s_{ij} and s^s_{ij} respectively.

10. The last correlation calculation to be performed over the box centered at (x_i, y_j) is to calculate the SCF with $\tau = 0$ and $s = 1$, the straight squared difference between each pair of spectra. The resulting averaged value is referred to as S^0_{ij}.

11. Finally, the above correlation calculations are performed for each spectrum in the cube serving as the base spectrum. This procedure yields values of the parameters $T^*_A, \Delta v, v_{LSR}, T_{int}, \sigma_{rms}, S, \tau, s, S^l, \tau^l, S^s, s^s$, and S^0 for every point (x_i, y_j). Maps of each of these parameters can be generated and in their analysis, the results of the SCF can be determined.

The SCF algorithm was written in the IDL software package to take advantage of the built-in functions, the multi-dimensional array processing, and graphics utilities. It would be easy to port the code to FORTRAN or C, for speed in processing, though the output would have to be analyzed using some graphics package, like PGPLOT or MONGO. The final code appears in the appendix as well as its necessary subroutines.

4. Understanding the SCF

In some senses the title of this section may offer more than it can produce; however, the following is an attempt to display some of the behaviors of the SCF and extract meaning from them.

The first step in interpreting the results of the SCF was to apply the algorithm to small cubes of completely artificially generated spectra. It is important to note that these do not represent the data from the simulations in any way. Rather, they are simple constructions of data cubes which are intended to display the behavior of the SCF with only one regulated aspect of the data cube changing.

The data cubes generated had 484 spectra arranged in a 22×22 array. These spectra were Gaussians and had various amplitudes, widths, and/or velocity offsets. Several different configurations were analyzed including random parameters, sudden jumps, linear gradients, and constant variations in the parameters. A few general statements about the outputs from the SCF can be made.

The correlation function S represents the similarity in shape between the surrounding lines. In a test where the spectra were all Gaussians with the same width, the value of S was 1 for the entire map, even for randomly distributed offsets and amplitudes. Since these all have the same shape, the correlation function accurately demonstrated the similarities of these lines.

The lag parameters τ and τ^l both represent the velocity by which the neighboring spectra are offset from the test spectra. For constant gradients (i.e. linearly changing values of v_{LSR}), this number is zero. This result is due to the fact that for any given test spectrum, the comparison spectra with higher velocities are balanced by those with lower velocities, so the mean shift is zero. Any deviations from this can be attributed to the discrete spacing of the grid points and the inability for the spectra to be shifted in order that a perfect match be obtained.

The lags are interesting because for smooth Gaussians, a velocity gradient of order m can be shown to have a corresponding lag field of order $m - 2$, implying a relationship similar to that of taking the second derivative. Ordinarily, recognizing the relationship would be a boon to calculations, but the relationship only holds true when the peak velocity is the only thing differing between two spectra. When their shapes and amplitudes differ, the correlation might be maximized with a lag that is not the one predicted by the second derivative of the velocity field.

The scaling factors s and s^s represent the similarities between the antenna temperatures of the spectra. When the lag is turned on, the value s represents the similarity between the integrated areas temperature of the spectra, with a value of one indicating equality. When the lag is off, the value, s^s, represents the mean, for all velocities, of the ratio between the antenna temperatures from the two spectra at given velocities. A value greater than one indicates that the base spectrum is consistently higher than comparison spectrum and vice versa.

The remaining correlation functions contain important information in what is held constant rather than what is allowed to vary. For example, S^l, the correlation with adjustable lag but not scaling of the spectra is helpful for comparing the similarities of line profiles. A high value will indicate that the spectra are similarly shaped and scaled though they may not necessarily be from regions with the same bulk velocity. On the other hand, the correlation function with zero lag but adjustable scaling, S^s, measures the similarity in velocity distribution at two positions without regard to the amplitude changes.

Finally the correlation function with the lag off, S^0, indicates simply the similarity of the two spectra in size, shape and velocity distribution. Strong correlations are tied to features that are large compared to the resolution of the SCF; thus changing the resolution will highlight different structures.

In this section of the paper, several examples of artificially generated data are interpreted. The relevant maps are displayed and others are summarized in the text. The best approach to interpreting these data may be simply to leap in and look at some examples.

4.1. Velocity Jump

The first example contains the results of the SCF algorithm when the spectra in a map differ from each other by a sudden jump in velocity. In this case, the jump in v_{LSR} is from 1 m/s for low values of y to -1 m/s for high values of y. Physically, this jump corresponds to the spectra in the bottom half of the graph having a red-shift and the spectra in the top half of the graph being blue-shifted. These are the spectra of two bodies of gas: one coming towards the observer and one retreating from the observer. The results are shown in Figure 2.

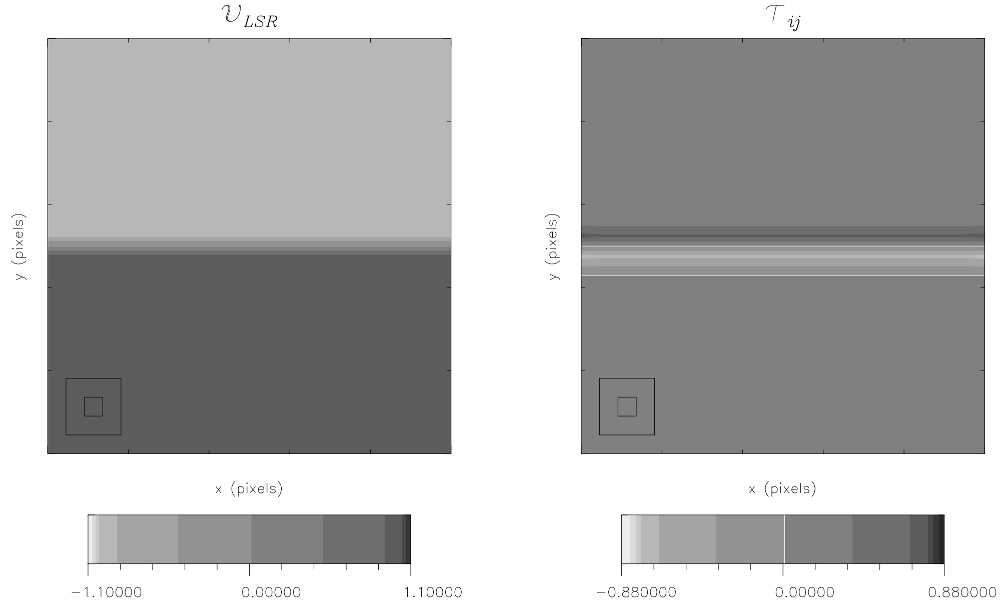

Fig. 2.— Maps of artificial data with a jump in v_{LSR}.

The figure on the left is a grey-scale representation of the fits to the v_{LSR} of the generated spectra. It depicts the sudden jump in v_{LSR} halfway up the data cube. The corresponding representation of the lag is displayed on the right. In all maps, the color bar extends for 10% beyond the actual values, which is why the extrema of the bar are at -1.1 and 1.1 respectively.

As expected, the map displays a jump when it crosses the boundary between red-shifted spectra and the blue-shifted spectra. The drop in the value of τ before the jump is because the spectra on the border that are red-shifted are being correlated with spectra in the blue-shifted region that have a negative v_{LSR} relative to the spectrum with which they are being correlated. There is a corresponding jump in the blue-shifted section where the spectra are correlated with spectra that have a greater value of v_{LSR} than the base spectrum for the correlation.

In all cases, the plots have a pair of nested boxes in the lower left-hand corner. These boxes represent the size of the pixel (inner box) and the size of the SCF $r \times r$ box where the correlations are conducted (outer box). In addition to these visual aids, an extra contour has been added at the zero level so that the examiner can easily distinguish between those regions with positive τ or v_{LSR} and those with negative values.

All other plots of the map yield what is expected. The SCF function is equal to 1 over the entire plot in cases when both scaling and lag (S_{ij}) and just lag (S^l_{ij}) are turned on. In the case where there is no lag turned on (S^s_{ij} and S^0_{ij}), the maps appear as shown in Figure 3.

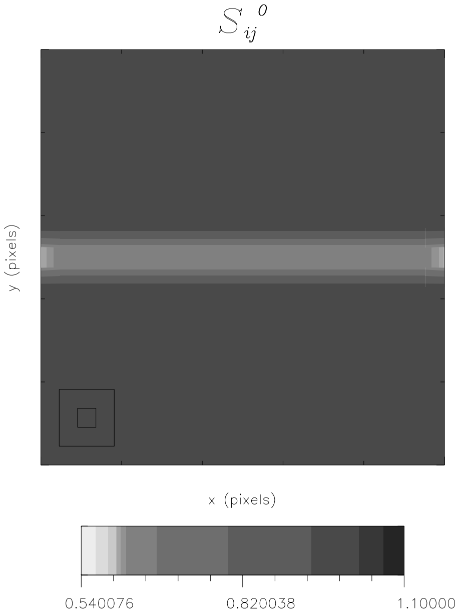

Fig. 3.— Map of S^0 in the case of a jump in v_{LSR}

The band in the center adequately represents the fact that the spectra in this region are not well correlated with their neighbors. The small boxes on the edge of even poorer correlation are the result of an edge effect. It is because of effects like these that the r pixels on the border of the map are discarded (here r is the resolution of the SCF).

Because the artificial spectra were all of the same height, the scaling factor, s is uniformly 1. The parameters from the fitting indicate that the routine fits the Gaussians well.

4.2. Amplitude Jump

The amplitude jump is almost exactly the same as the velocity jump, except that the jump is in the height of the Gaussian in the model spectra. The jump is from high amplitude at low y to low amplitude at high y, representing two different regions of gas, one with a higher peak antenna temperature than the other (See Figure 4).

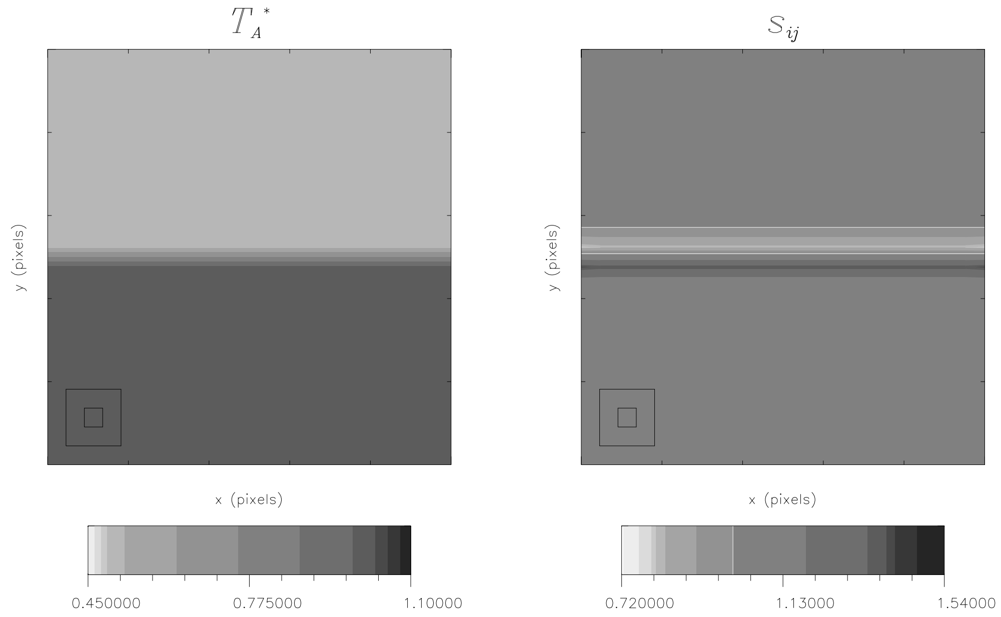

Fig. 4.— Maps of artificial data with a jump in T_A^*.

Again, there is a jump before and a dip after the change in the amplitude. The rise before the change may be a bit counter-intuitive; however, the way that s is defined is the amount by which the neighboring spectra are scaled up. Thus, the low spectra being scaled up to have the same form as the high spectra is represented by the jump and the high spectra being scaled down is represented by the dip. A border contour is plotted here as well, this time dividing regions with $s < 1$ from those with $s > 1$. Those with $s < 1$ indicate that the neighboring spectra had to be scaled down, so the spectrum here has, on average, a smaller integrated area than its neighbors.

The other plots generated by the SCF algorithm are similar to those generated in the case of the velocity jump. The SCF plots without scaling turned on are identical to those in the velocity jump without lag turned on (see Figure 3 and comments pertaining thereto). Similarly, those plots of the SCF with scaling turned on are uniformly equal to one (with an error of 1 part in 10^7 due to fluctuations in the floating point calculations used on the computer).

4.3. Width Jump

While differences in lag and scale can be compensated for by the parameters τ and s respectively, differences in shape can not be so corrected; thus the SCF can be used to measure the differences in shapes between lines. The result is that the SCF defines correlated spectra to be those with similar shape. A change in width is tantamount to a change in shape.

In this set of data, the jump is from large widths for low y to small widths for large y. The relevant plots for the case of changing widths are found in Figure 5.

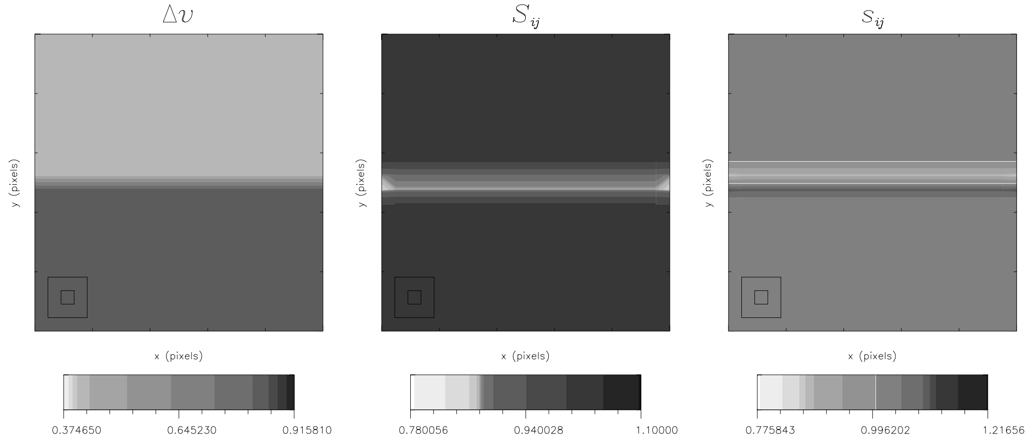

Fig. 5.— Relevant plots for a jump in Gaussian width

The figure on the left represents the fit to the Gaussians, depicting the jump from large width to small width as y increases. The second plot, representing the SCF, shows how the width changing affects the data because the shapes are different along the interface between wide and narrow lines. There is less correlation between the spectra at that point. Finally, the last plot represents the value of s changing, which stands to reason because the SCF algorithm uses a changing scale factor to compensate for the differing widths.

4.4. Resolution

One of the input parameters that can be changed is the resolution, r, in order to regulate the size of the box over which the SCF parameters are evaluated. Again, the case of a changing v_{LSR} is used, but in this case there are two small clumps of gas with strong red-shifts against background at rest relative to the observer. Like resolution in optics, the higher the value of r that is used, the more difficult it is to distinguish between two distinct features and to deduce their physical extent. The plots in Figure 6 illustrate this effect.

For the figure with resolution 13, it is impossible to determine whether the gas distribution has a large region of spectra that are not perfectly correlated or a smaller region that is pulling the averages down. For this reason, the smaller resolution can be used to pick out these details.

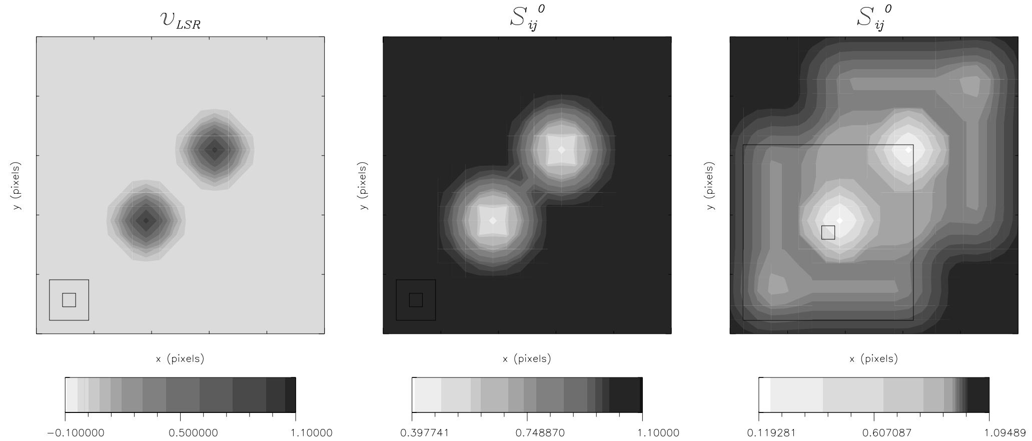

Fig. 6.— Relevant plots to display changing resolution. From left to right, the plots are of v_{LSR}, the parameter S^0_{ij} at resolution 3 and then at resolution 13.

Sometimes it can be advantageous to use low resolution on purpose. The correlations between gas clumps at small scales can be detected with high resolutions, but these features tend to mask the larger scale correlations present in a data set. To illustrate this, a data set has been generated with small-scale and large-scale structure. The relevant plots appear in Figure 7.

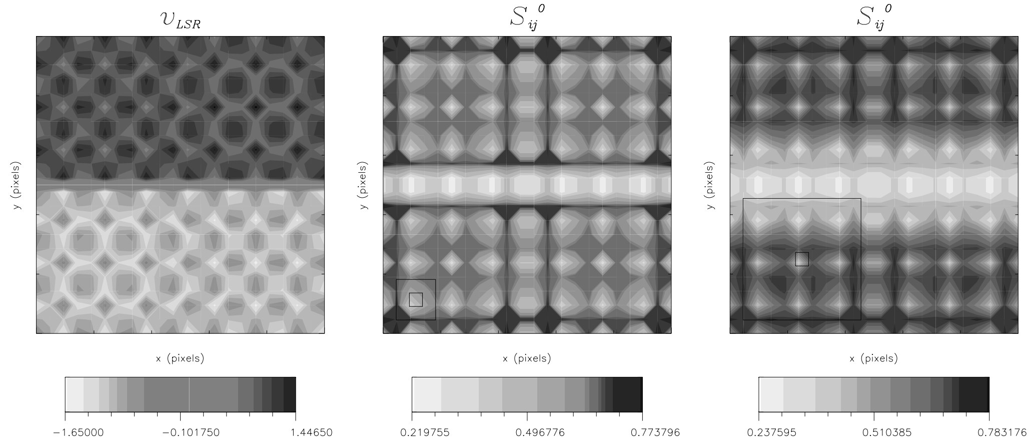

Fig. 7.— Relevant plots to display correlations on different scales. From left to right, the plots are of v_{LSR}, and the correlation measure S^0 at resolution 3 and then at resolution 9.

The velocity field is one with a small-scale sinusoidal modulation, on order of a pixel size, and a large jump at the center from blue-shifted velocity to red-shifted velocity. The jump is not readily distinguishable in the correlation measure at low resolution. There is a feature there; however, it cannot be discerned from the smaller scale correlations. On the other hand, if the resolution is increased, the large-scale correlation of the gas on either side of the jump becomes more apparent. The increased resolution serves to highlight this feature. The jump will become more dominant as the resolution increases, washing out the small-scale modulation in the infinite limit. This observation implies that the size scale of the clumps of gas on either side of the jump is infinite compared to that of the modulations. In examining the fashion in which the jump has been set up, this stands to reason because the jump looks like a small section of the border between two differently moving clouds of gas.

4.5. Bringing it all together

The above differences are rather easy to understand individually; however, when all three basic effects are present in the same cloud, the interpretation is more difficult. In order to better understand such amalgamations of effects, the compensating factors s and τ can be turned on and off at will.

First, the advantages of normalization immediately become apparent. We can examine a clump in a cloud and compare its values for the various flavors of the SCF. With both scaling and lag turned on, the value of S represents the best correlation one can get between two spectra. This value serves as a basis for comparison. Then, we can turn off lag ($\tau \equiv 0$) and scaling ($s \equiv 1$) respectively and compare the resulting S^s_{ij} and S^l_{ij}. If one of these is significantly higher than the other, we can assert that the gas has similar velocity distributions ($S^s > S^l$) or that it is emitting similar amounts of radiation, but not moving coherently ($S^l > S^s$). If the SCF with both lag and scaling turned off is comparable to those values with it on, we can assert that the gas within one resolution box of the central spectrum has either a roughly uniform or a completely chaotic nature. The next fashion in which the results of the SCF can be compared is along the lines of the actual value of the SCF. The function, after all, is a measure of the correlation between spectra and with the normalization factor included, different parts of the map can be compared. Additionally, maps can be compared with each other. Regions with correlation close to one are similar in shape to the spectra within the resolution box. Similarly, the calculated values of the other SCF variations, S^s and S^l, indicate similarities in intensity and velocity distribution over the same areas.

In addition to mapping the SCF and its subsidiary variations, the parameters s and τ that maximize the correlation for each point on the map are also calculated. Their maps can be generated to give further information about the clouds.

The easiest aspect of these maps to consider is the magnitude of the lag parameter τ. The parameter takes on large magnitude values when calculated along the border between two differently moving regions of gas. It is similar to the velocity gradient of the cloud, but by altering the resolution of the SCF, the structure of a velocity shift can be interpreted. For example, the resolution at which the jump appears the most prominent is the size scale of the physical jump. In order to assist in the interpretations, plots of $|\tau|$ and $|\tau^l|$ are made as well as color plots of the variables using a special color table to highlight these changes.

The maps of S^s,S^l, and S^0 are the actual keys to understanding the structures of clouds. Unfortunately, the discussion of these maps of the data in concert is rather useless with the simple data cubes that have been generated. The discussion of these factors is, therefore, postponed until the analysis of real data.

Finally, it must be noted that there tends to be a stronger correlation between data with high signal to noise than those with poor signal to noise. This is an artifact of the noise primarily because those signals with noise that is comparable to the strength of the line will never be able to line up as large a fractional area under the subtracted curve as will those with larger signal to noise. This bias can be partially removed with smoothing. The signal to noise in the data is an important factor in the reliability of the results.

With these realizations about the SCF and its behavior, the data received from actual clouds can be understood.

5. Analysis of Heiles Cloud 2

Observers at FCRAO observed Heiles Cloud 2 in 1996 in the C^{18}O line at 109.782168 GHz. The resulting data cube consisted of 4800 spectra arranged in a grid of 50 × 96 pixels on the sky. The grid covered 1.38^m in right ascension and .667° in declination, centered on the cloud. The spectra themselves have 256 channels of velocity running from -0.35 km/s to 12.45 km/s, with 0.05 $km/s/channel$. The peak emission from the cloud is at about +6 km/s.

5.1. Signal to Noise

The typical signal to noise ratio for the map had a values of about 3. In order to eliminate problems resulting from the noise in the data, the spectra were smoothed to reduce noise. A Gaussian was fit to each spectrum and then each point in the spectrum was replaced with the smoothed value which was calculated in the following fashion:

$$\overline{T_A(v_i)} = \frac{1}{N}\sum_{j=0}^{N-1} T_A\left(v_i + \left[j - \frac{N}{2}\right] dv\right) \tag{9}$$

Here, dv represents the channel width and N was chosen to be the number of velocity abscissae in one HWHM. This smoothing width was chosen because it was the value which maximized the correlation function for a given data set, thereby representing the smoothing that resulted in the best possible correlations. This number is fairly universal for the data sets examined.

A sample spectrum and the result of the smoothing are plotted in Figure 8 for comparison.

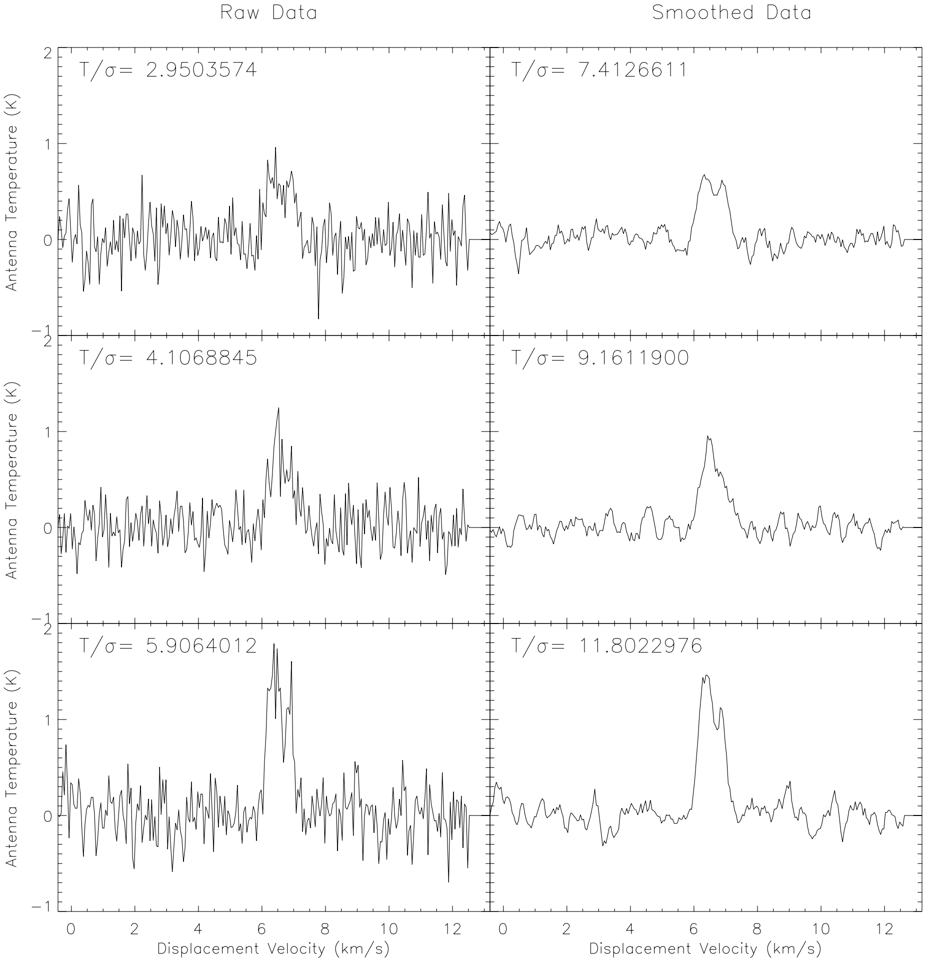

Fig. 8.— Demonstration of smoothing algorithm. Unsmoothed data on the left and smoothed data on the right. The number in each box represents the old and new signal to noise respectively.

In order to better estimate the effects of the noise in the spectrum, a Monte Carlo simulation was performed. In this simulation, a cube of perfect data was created. This cube consisted of 9 spectra, arranged in a 3×3 grid, each of which was a perfect Gaussian with a FWHM of 1.7 km/s and a uniform height. To each of these spectra, normally distributed noise was added in a fashion that set the signal to noise at a specific value. This was repeated 10 times to create 10 different cubes all with the same signal to noise. Each of these noisy cubes was processed with the SCF and the correlation functions were plotted as a function of signal to noise. Because each of these functions should have a value of 1 in the case of no noise, the value that the simulation yields should be the factor by which the correlation function is in error for a given signal to noise value. By reducing the signal to noise using the smoothing routine, better values of the SCF can be estimated. Running longer and more detailed simulations over a larger range of cases will produce a relation between the signal to noise and the error in the correlation function. Examination of the plot indicates that spectra with a smoothed signal to noise of less than 7 will result in inaccurate correlation results. Consequently, the threshold value for original signal to noise is set at 3. A plot of the behavior of the correlation functions as a function of signal to noise appears in Figure 9.

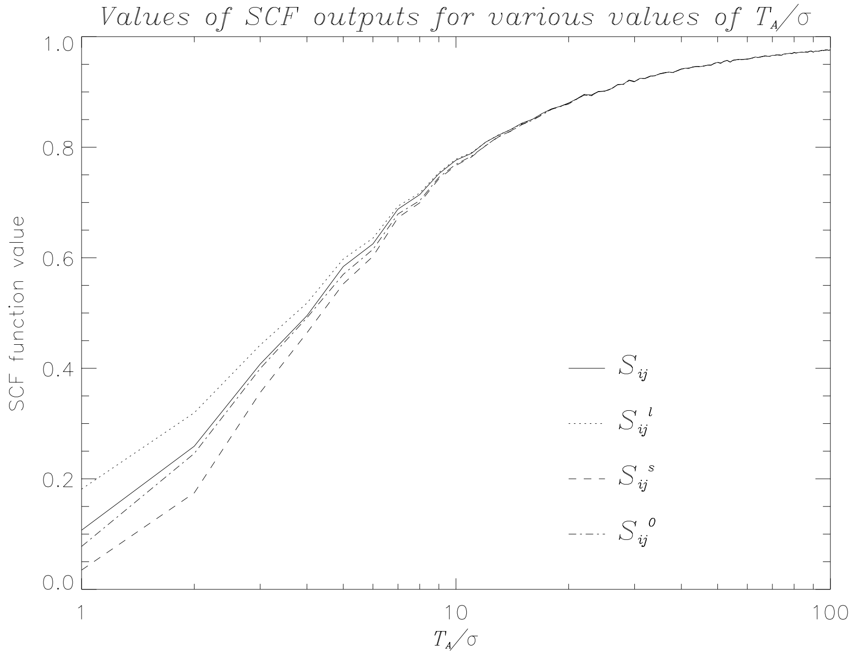

Fig. 9.— Effects of Noise on the values of the Correlation Functions

5.2. Maps of Heiles Cloud 2

The cloud maps can be analyzed in two ways: visually and statistically. This section deals with the former. The SCF algorithm processed the entire cube with a resolution box of 5 pixels ($r = 5$). All spectra with signal to noise less than 3 were rejected. Three FWHMs were considered in the calculations of the spectra ($q = 3$). With these parameters for the SCF, the maps of the cube can be generated, the most edifying of which appear in Figures 10 and 11.

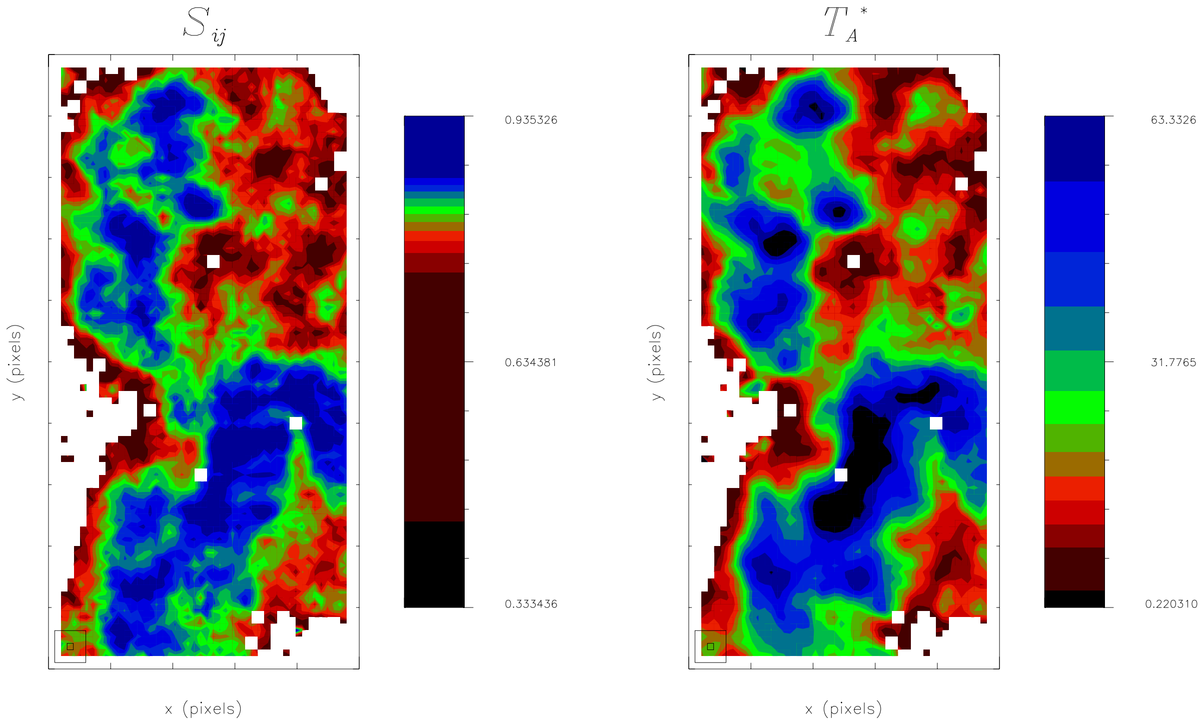

Fig. 10.— Maps of Heiles Cloud 2 Using the SCF

The figures include the correlation function with both lag and scaling turned on (S_{ij}), and the antenna temperature (T_A) in Figure 10. Blank pixels are the locations of rejected spectra and the 2 pixel wide border around the map is rejected because of the edge effects discussed in §3. The maps are in a grey-scale designed to display a dynamic range of the data. As a result, different colors correspond to different values in each plot. In order to aid in interpreting these maps, color bars are placed with each map. As mentioned previously, the scales run from 10% above the highest value to 10% below the lowest value.

The most interesting thing to note about these plots is that, while the basic structure is similar to that of the temperature map, there are distinct differences, especially near the maxima of each of the maps. To further illustrate this point, more maps from the SCF algorithm were generated. Figure 11 depicts the correlation function with only lag on (S^l_{ij}) and the correlation function with only scaling turned on (S^s_{ij}).

This map shows the differences between the regions of similar emission (S^l_{ij}) and those of similar velocity distributions (S^s_{ij}). There appear to be regions which are similar in one map but dissimilar in the other. By locating the regions which are of high correlation in all maps, the most similar features can be highlighted. Interpretation of these maps, however, is still in the rudimentary stages and results are

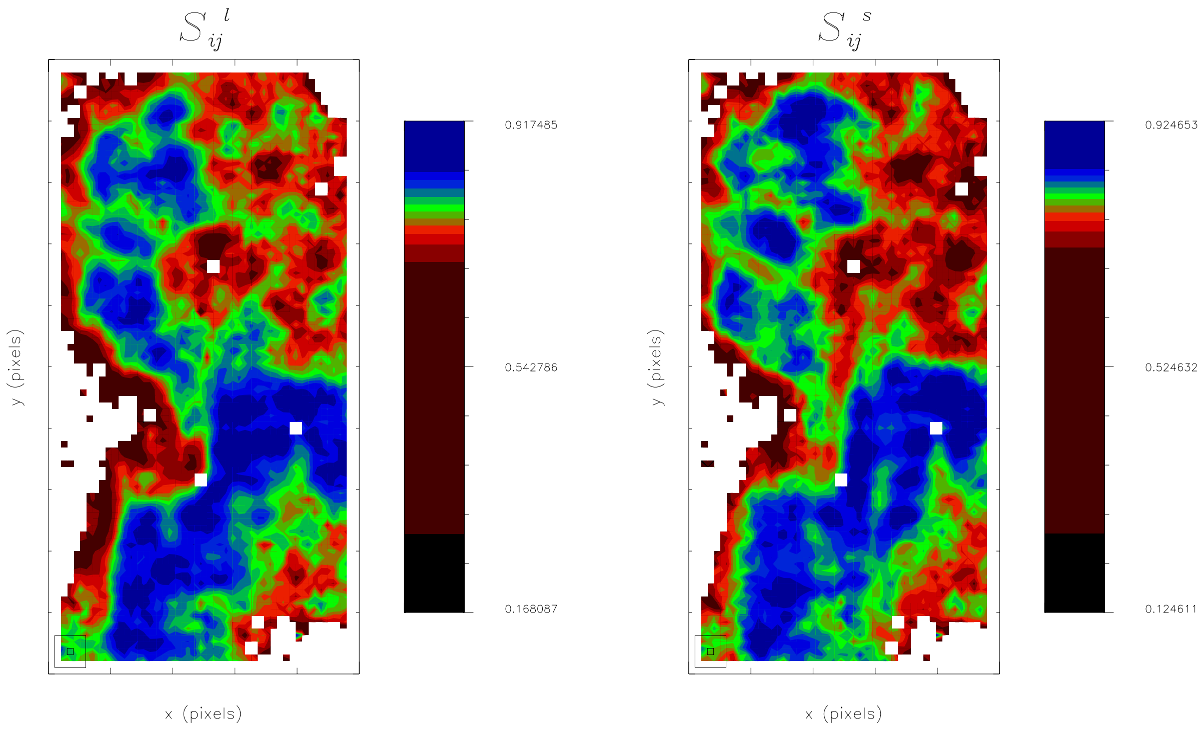

Fig. 11.— More maps of Heiles Cloud 2 using the SCF

anticipated soon. A complete set of maps appears in the Appendix.

The above maps can be analyzed by the human eye and high correlations can be noted, but the intent of the paper was to develop quantifiable measurements of cloud correlation. As a result, statistical methods were developed to aid in the analysis.

6. Statistical Analysis of SCF maps

The simplest analysis of the data is to consider the correlation functions over the map as sets of values. These values were binned with appropriately sized intervals and a histogram was generated. Statistical moments of the data were also calculated and these numbers can be used to compare maps. Data from computer simulations could also be processed from the SCF and the moments and histograms of these maps could also be calculated. However, at the time of writing, no such data cubes were available. As a result, another data set was generated by randomizing the positions of the spectra in the Heiles Cloud 2 data set. This process placed spectra that were originally from different parts of the map, having different properties, close to each other. Additionally, the same number of spectra were rejected and there was an identical distribution of signal to noise. Naturally, the correlation functions should have lower values for such a map than for the data cube of real observations. The most vivid illustration of this difference was found in the correlation functions without lag or scaling (S_{ij}^0) and these histograms appear in Figure 12. A complete set of histograms for both sets of data appear in the Appendix.

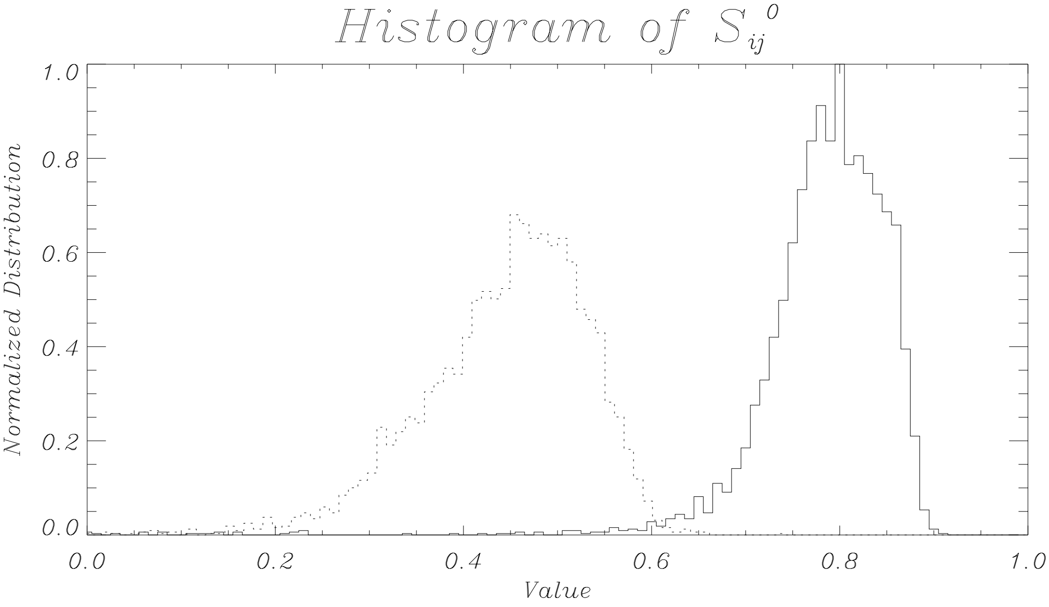

Fig. 12.— Distributions for the Heiles Cloud 2 Maps. The data for the regular map is shown in a solid line, and the data for the randomized map appear as a dashed line.

It is rather easy to see that the distribution of the real data has a much higher mean correlation and lacks the large tail seen in the histogram of the randomized data. These two facts can be seen as a result that the randomized data, on average, has a significantly lower correlation that the real data, exactly what is expected. This observation serves as a demonstration that the SCF can distinguish between spectra that are similar being next to each other and spectra that are obviously different, an aspect that appears in some of the simulations. In order to summarize the statistical analysis, the moments of the various maps are listed in Table 1.

Perusal of this table quickly indicates that, on average, the normal data cube has a higher mean correlation and a larger skewness than the random data representing the fact that the spectra are better

	Real Data				Randomized Data			
Parameter	Mean	Deviation	Skewness	Kurtosis	Mean	Deviation	Skewness	Kurtosis
S_{ij}	0.84	0.057	-2.1	10.	0.72	0.064	-1.4	5.1
τ_{ij}	0.00024	0.028	-0.12	3.3	-0.0018	0.23	0.25	0.040
S^l_{ij}	0.89	0.0046	-2.9	18.	0.55	0.090	-1.1	1.6
τ^l_{ij}	0.00048	0.0090	0.065	9.5	-0.0033	0.24	0.24	0.66
S^s_{ij}	0.82	0.071	-4.0	29.	0.55	0.090	-0.97	2.0
S^0_{ij}	0.78	0.020	-4.5	35.	0.44	0.087	-1.0	2.3

Table 1: Table of Statistical Moments for Heiles Cloud 2 Data

correlated and that the majority of the values are close to the mean with a tail off to smaller values. For the randomized data, the larger deviation and

lower skewness represents a wider distribution with a less drastic edge above the mean. The negative value of skewness adequately represents the absence of a tail of values towards higher correlations. The mean of the S_{ij} function for the randomized data is close to that for the normal map, indicating that the functions are roughly of similar shape over the entire map. Finally, it can be noted that the deviation of the distributions of v_{LSR} is 0.41, which is similar to the deviation of the values of τ for the randomized data. The deviations of τ from the regular map were much lower. This observation indicates that the actual map contains regions that exhibit similar velocity distributions, and these distributions are specific to areas of the cloud. By comparing these moments, the two data sets can be distinguished and their similarity judged.

7. Discussions and Conclusions

From the above results, we can conclude that the Spectral Correlation Function can be used as a tool to distinguish between data sets that have similar spectra grouped together and those that do not. Using the moments technique outlined in the final section, the goal of quantifying the similarity of neighboring spectra has been realized. As yet, these numbers indicate a difference between the two data sets. Interpreting the meaning of these moments beyond mere dichotomization between random and real data sets has yet to be realized and is a goal for the future.

In addition to using moments to analyze the histograms, more analysis of the maps could yield interesting results. The maps beg the question of whether there is any physical meaning contained in the distinct Regions of Correlation (ROCs) that appear in the maps. The fact that different ROCs are highlighted at different resolutions and with different correlation functions may give some clue to the meaning of these ROCs. This analysis will have to be postponed until future work.

In addition to these directions, future work will also attempt to compare the results of the SCF algorithm between different clouds, as well as to analyze data from simulations. Other types of Spectral Correlation Functions will be developed, hopefully, to exhibit different kinds of correlations and to explain the results that have be calculated to date.

The Spectral Correlation Function can be used to map out data sets in terms of the correlation of their constituent spectra, and statistical moments of these maps can distinguish between data sets.

Acknowledgments

I would like to thank several parties for their contributions to this paper. Primarily, I wish to thank Alyssa Goodman, my advisor, and Dave Wilner, my mentor, for their unending support and direction on this project. Additionally, I would like to thank Marc Heyer of FCRAO for the use of the Heiles Cloud 2 data set before its official publication. Finally, I owe many thanks to Jonathan Williams for his help and orientation with IDL and insight with the project.

REFERENCES

Dyson, J.E., and Williams, D. A. 1980, The Physics of the Interstellar Medium. (Manchester University Press : Manchester), 166ff.

Gammie, C.F., Ostriker, E.C. 1996, Ap. J., 466, 814.

Goodman, A.A. 1997, `http://cfa-www.harvard.edu/~agoodman/scf/scf.pdf`.

Falgarone, E., *et. al* 1994, Ap.J., 436,728.

Fuller, G.A., Myers, P.C. 1992, Ap. J., 384, 523.

Larson, R.B. 1981, MNRAS, 194, 809.

Myers, P.C. 1987 in Interstellar Processes, ed. D.J. Hollenbeck and H.A. Thronson, Jr. (Dordrecht: Reidel), 71ff.

Passot, T., Vàzquez-Semadeni, E., and Pouquet, A., 1995, Ap. J., 455, 536.

Pound, M.W., Goodman, A.A. 1997, Ap.J., 482, 334.

Vàzquez-Semadeni, E. 1996, Ap. J., 473, 881.