Chapter 2

The SCF Algorithm

The focus of the SCF project is to develop a function that represents the correlation between two spectra and generate a computer algorithm that efficiently calculates this correlation. The function is a type of two-dimensional autocorrelation function with additional parameters included to investigate the nature of the correlations over an entire spectrum. This task is subdivided into three distinct parts, each of which is described in the following sections. First, a meaningful correlation function must be defined. Next, such a function must be developed into a working computer algorithm. Finally, the presence of noise in the data must be addressed in order that a meaningful comparison between spectral maps can be made.

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

2.1 Mathematical Development of the SCF

2.1.1 The Deviation Function

To measure the correlation between two spectra, each spectrum is viewed as a function in velocity space: $T_A(v)$. A new function, called the deviation function, is defined based on these two functions:

$$[\delta_{1,0}(v)]^2 = [T^*_{A,1}(v) - T^*_{A,0}(v)]^2 \tag{2.1}$$

The deviation function is in units of antenna temperature squared $[K^2]$ and is identically zero for identical spectra. In order to develop a better sense for what the deviation function means, it is reduced to a single number by integrating in velocity space, resulting in a single number with units of $K^2 \cdot km/s$. This number is referred to as the deviation scalar or simply the deviation.

$$D(T_1, T_0) \equiv \int [\delta_{1,0}(v)]^2 dv = \int [sT^*_{A,1}(v - \ell) - T^*_{A,0}(v)]^2 dv \tag{2.2}$$

The free parameters s and ℓ have been introduced so that the deviation can be minimized between the two spectrum, allowing for spectra with similarities in shape and differences in amplitude or velocity offset to be recognized as having some correlation. The

introduction of these parameters is what distinguishes the SCF method from the other analysis routines discussed above. These parameters recognize more kinds of similarities as significant than do the three dimensional analyses or the correlation function methods.

The deviation is also examined with $s = 1$ and/or $\ell = 0$ to understand the differences between the two spectra. The values of these parameters offer hints to the relations between the two spectra. For example, if the value of ℓ 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.

The deviation function is non-negative by its definition so a deviation of zero is its lower bound resulting from the integral of the zero function. To determine the maximum value of the deviation, the functional definition 2.2 is expanded:

$$D(T_1, T_0) = s^2 \int [T^*_{A,1}(v - \ell)]^2 dv - 2s \int T^*_{A,1}(v - \ell) T^*_{A,0}(v) dv + \int [T^*_{A,0}(v)]^2 dv \qquad (2.3)$$

In the above equation, the cross term represents the correlation between the two functions. If the two spectra are radically dissimilar, this cross term will go to zero. For example, consider two lines radically separated in velocity space so that the near-zero wings of one function are multiplied by the peak of the other and vice versa. In the limit of poor correlation, the integral of the cross term will be zero and the maximum value of the deviation is simply the exterior terms:

$$D(T_1, T_0)_{max} = s^2 \int [T^*_{A,1}(v - \ell)]^2 dv + \int [T^*_{A,0}(v)]^2 dv \qquad (2.4)$$

This understanding of the range of the deviation function becomes essential in normalizing the results of the functions.

2.1.2 Normalization

In order to aid in interpreting the results of the deviation calculation, the value of the calculation can be normalized to the unit interval. The desired normalization is to have a value of 1 indicating perfect correlation and the value of 0 indicating the minimum correlation in the following fashion, thereby defining the Spectral Correlation Function $S(T_1, T_0)$:

$$S(T_1, T_0) \equiv 1 - \sqrt{\frac{D(T_1, T_0)}{s^2 \int T_1^2(v) dv + \int T_0^2(v) dv}} \qquad (2.5)$$

The normalization value in the denominator is chosen because it represents the maximum value of the deviation in the absence of absorption (c.f. Equation 2.4). Absorption in observed spectra only occurs in a few cases and the result is that the antenna temperature of a spectrum becomes negative. Thus, the cross term in Equation 2.3 will be positive if there is a large contribution due to absorption in one of the spectra. In this case, the final normalized deviation will be negative. This is because $\int [\delta(v)]^2 dv > s^2 \int T_1^2(v) dv + \int T_0^2(v) dv$. These effects will result in $S < 0$ and, because the scale factor s is derived using similar integrals, its value will also drop below zero.

2.2 Efficient Optimization of the Deviation

The SCF algorithm will necessitate the calculation of large numbers of deviation functions; hence these calculations will need to be done efficiently and accurately. Minimization of a function in parameter space is the subject of much work in the field of numerical analysis and there exist many routines to use in such minimization problems. Unfortunately, the deviation, viewed as a function of s and ℓ is more difficult to analyze than most functions to which these methods are applied. The main reasons for this difficulty lie in the parameter ℓ. These difficulties can be grouped into three major problems:

1. The value of $D(T_1, T_0)$ can only be evaluated for values of ℓ such that $T_A^*(v-\ell)$ is defined by the original spectrum. In other words if $v' = v - \ell$, 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 $\frac{\partial}{\partial \ell} D(T_1, T_0)$. Any accurate partial derivative of $D(T_1, T_0)$ with respect to ℓ requires evaluating $T_A^*(v-\ell)$ at several points. For noisy functions, the number of points evaluated must be quite high in order for the partial derivative to be of any use to a numerical optimization routine.

3. The function $T_A^*(v)$ has many local extrema because of noise and thus any minimization routines, in seeking a minimum with respect to ℓ, 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 global minimum. However, for typical spectra, this would require a large number of iterations, far more than most programs are intended to use.

To develop an understanding of what will be required from a minimization routine, Figure 2.1 displays the deviation in $s-\ell$ parameter space. The two large humps in this figure represent the increase in deviation that occurs when both peaks in the spectra are present in the calculation window, but are not aligned so that the cancelation can occur. The drop off that occurs past these two peaks is because the spectra are not of infinite span in velocity space and the peaks migrate off the portions of the spectra which are being examined. By considering lines of constant ℓ, one important feature is noticed: the variations in s result in a parabolic variation of the deviation. This type of relationship is capitalized upon in the development of the minimization routine because the minimum, for a given value of ℓ, is easily calculated. By viewing the function as a paraboloid in s, a simple derivative will produce the appropriate minimizing value of s.

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

$$\Rightarrow s_{min} = \frac{\int T_{A,1}^*(v-\ell) T_{A,0}^*(v) dv}{\int [T_{A,1}^*(v-\ell)]^2 dv} \tag{2.6}$$

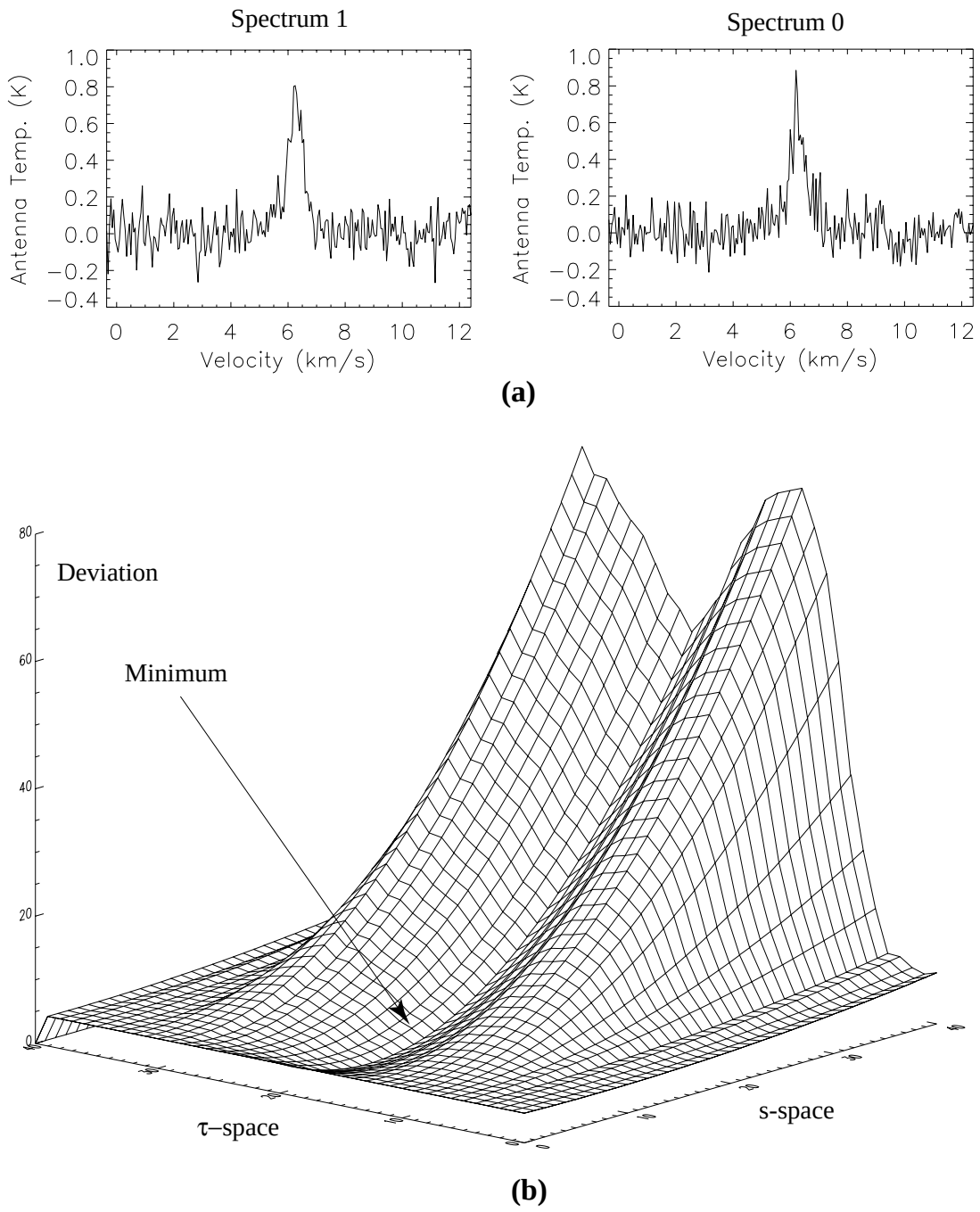

Figure 2.1: $D(T_1, T_0)$ viewed in parameter space. (a) The input spectra. (b) The parameter space.

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 ℓ. The presence of a single minimum in s-space is fortunate because it allows for the minimization of the deviation to be done serially. The process is to find the minimum value of the deviation for any given value of

ℓ by adjusting s and then selecting the minimum of all these values. The resulting value is guaranteed to be the minimum of the deviation because it is the minimum of the complete set of local minima. Computing the minimization in the opposite order does not necessarily guarantee that the set of local minima is complete and thus the global minimum might not lie among these. In addition, the fewer times the deviation is calculated, the better; for calculations with changing ℓ are very time-consuming.

2.2.1 Numerical Integration

The evaluation of the deviation, $D(T_1, T_0)$, requires the computation of the integrals in equation 2.3. In addition, minimizing s requires the computation of the same integrals. There are many numerical integration routines available on different computer platforms. However, efficiency is a priority in designing the algorithm and the simplest numerical method is 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: $v_{i+1} - v_i = \ constant \ \equiv \delta v$ for all i. In order to integrate the spectrum over the velocity range, the values of the function are summed and the resulting total is multiplied by δv, corresponding to the rectangular estimate of the area under a curve.

$$\int f(v)dv \simeq \left(\sum_i f(v_i)\right) \cdot \delta v \tag{2.7}$$

This approximation is excellent, so long as the sampling of the data, δv, is smaller than the scale over which the function $f(v)$ varies appreciably. For Gaussians, an integration accurate to 1 part in 10^6 requires but 2 samples per half width. Unfortunately, most spectral data are 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 ideally be zero and can be used as the error in the antenna temperature values. 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 due to noise in the calculation is:

$$\partial f(v) = \sqrt{T_j^2(v)\sigma_i^2 + T_i^2(v)\sigma_j^2} \Longrightarrow \tag{2.8}$$

$$\partial g = \int \partial f(v)dv \simeq \left(\sum_i \partial f(v_i)\right) \cdot \delta v \tag{2.9}$$

In the above equation, a "∂" 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. The problems with correlation because of noise in the spectra will be addressed in the next section.

2.3 The Effect of Noise on the SCF

The presence of noise in the spectra interferes with the correlations between two spectra in a significant fashion. The reason for this is that noise changes the shape of the two spectra, and while the actual signal may be well correlated, the addition of noise will prevent the spectra from having similar shapes. There is therefore a bias in the SCF values favoring pairs of spectra with a high ratio of signal to noise.

In order to better estimate the effects of this signal to noise bias, a numerical simulation was performed. In this simulation, two perfect Gaussian spectra were compared using the spectral correlation function. The spectra had a FWHM value 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 100 times to create a large body of spectra with the same signal to noise. 100 pairs of these noisy spectra were 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.

In order to estimate the effects of the parameters s and ℓ on the behavior of the data, the spectra were compared using four different correlation functions with various combinations of the parameters. These functions are summarized in Table 2.1.

Function Name	s	ℓ	Spectral Property Examined
S_{ij}	Float	Float	Compares shapes of spectra
S_{ij}^{ℓ}	1	Float	Emphasizes similarity in shape and amplitude
S_{ij}^{s}	Float	0	Highlights similarity in velocity offset and shape
S_{ij}^{0}	1	0	Measures similarity in all properties

Table 2.1: Summary of the Correlation Functions Used in SCF analysis.

The behavior of these four correlation functions is plotted as a function of signal to noise in Figure 2.2. At low signal to noise values, the correlation functions which do not use a calculated value of s tend toward higher values than do those with the value given in Equation 2.6. This aberration is the result of the errors incurred by the numerical integration routines discussed previously. The noise in the spectra artificially reduces s factor in the calculation. This is because the squared spectrum in the denominator of Equation 2.6 is the square of a spectrum, allowing part of the noise to reinforce itself. The product of the noise from two different spectra has no such reinforcement. A more detailed discussion of the arguments behind this reasoning appears in Section 2.3.1.

2.3.1 The Seljak Correction

The signal to noise bias discussed in Section 2.3 must be circumvented in order that meaningful values of the SCF be calculated. In order to correct for these difficulties, Uros Seljak of the Harvard-Smithsonian Center for Astrophysics has proposed the following correction.

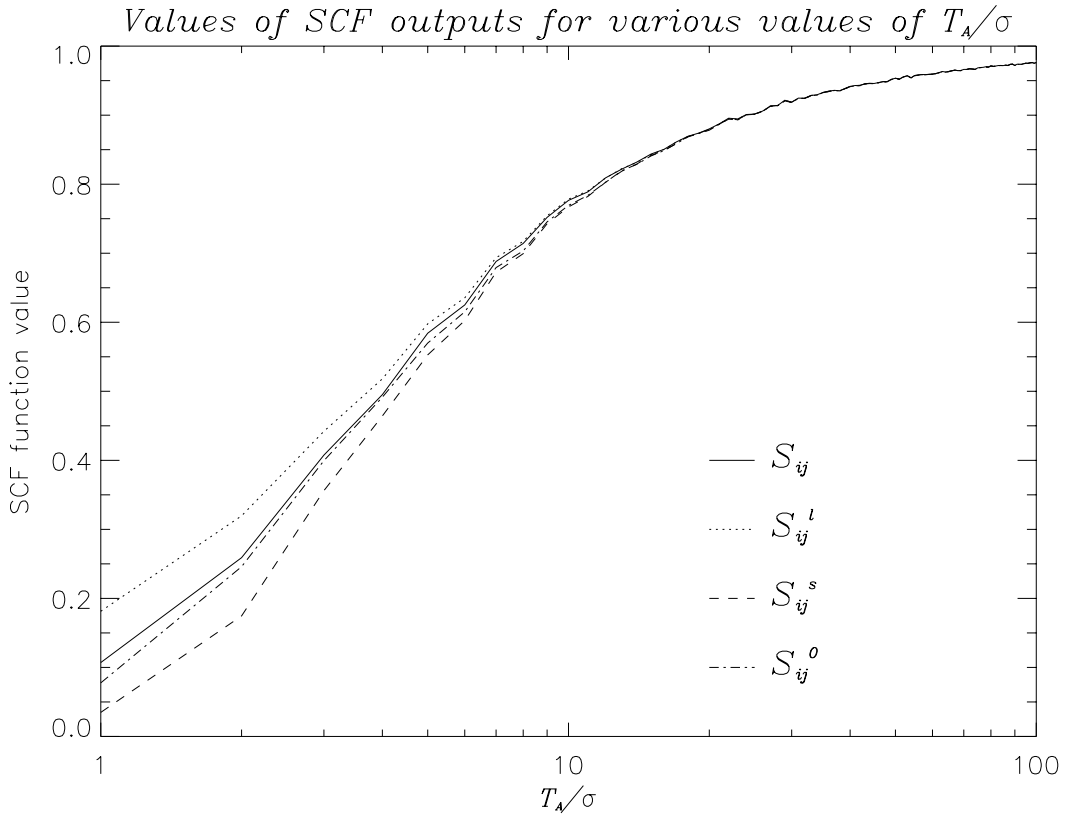

Figure 2.2: The behavior of the four correlation functions under the influence of different values of signal to noise.

Every spectrum can be considered as the sum of two different functions, a signal function and a noise function:

$$T_A(v) = S(v) + \mu(v) \tag{2.10}$$

Hence the deviation function can be expressed in terms of these two individual functions:

$$D(T_1(v), T_0(v)) = \int s^2\ T_1^2(v) - 2s\ T_1(v)T_0(v) + T_0^2(v)dv$$

$$= \int \left(s^2(S_1^2 + 2\mu_1 S_1 + \mu_1^2) - 2s(S_1 S_0 + \mu_1 S_0 + \mu_0 S_1 + \mu_1\mu_0) + S_0^2 + 2S_0\mu_0 + \mu_0^2 \right) dv \tag{2.11}$$

The principal insight in the Seljak correction is recognizing the individual terms of the integral in terms of their average values. For example, we can use the definition of the average value of μ_1^2 to evaluate the terms in the above integral:

$$\left\langle \mu_1^2 \right\rangle \equiv \frac{\int \mu_1^2(v)dv}{N_{pixels} \cdot \delta v_0} \Longrightarrow \int \mu_1^2(v)dv = \left\langle \mu_1^2 \right\rangle N_{pixels} \cdot \delta v_0 \tag{2.12}$$

In the above equation, δv_0 is the spacing between channels in the spectrum and $(N_{pixels} \cdot \delta v_0)$ is the range over which the noise is averaged.

The character of the noise in the spectra is assumed to be normally distributed noise with a mean of zero and a measured root-mean-squared deviation of σ. As a result of these features and the assumption that the noise and signal are statistically uncorrelated, the product $\langle \mu_i S_j \rangle$ or $\langle \mu_1 \mu_0 \rangle$ can be broken up into terms involving $\langle \mu_i \rangle$ multiplied by another term. According to our assumptions about the character of the noise, these terms are zero. Thus, equation 2.11 can hypothetically (see below) be reduced to:

$$D(T_1(v), T_0(v)) = \int s^2(S_1^2 + \mu_1^2) - 2sS_1S_0 + S_0^2 + \mu_0^2 \; dv \tag{2.13}$$

This reduction is accomplished by setting all terms that are linear in μ_i equal to zero using the argument given above. As given in equation 2.12, the values of $\int \mu_i^2 dv$ can also be calculated using the measured root-mean-squared value of the noise. Thus, reducing the calculation to only correlations between signal gives the corrected deviation:

$$D'(T_1(v), T_0(v)) = D(T_1(v), T_0(v)) - s^2 N_{pixels} \delta v_0 \sigma_1^2 - N_{pixels} \delta v_0 \sigma_0^2 \tag{2.14}$$

While this correction holds in the statistical limit, neglecting the terms that are linear in μ_i can be dangerous. Examining the rms deviation of these terms shows that the amount by which they deviate for a given signal, is larger than the correction applied because of the noise alone. Treating the signals as Gaussians the rms deviation of one of these terms is given by:

$$\Delta \left(\int \mu_i S_j dv \right) \simeq \sqrt{N_{pixels} \delta v_0 \sigma_i^2 A_j^2 \eta_j \sqrt{\frac{\pi}{2}}} \tag{2.15}$$

Here, η_j is the width of the Gaussian. The magnitude of these deviations are comparable to the calculated values for $\langle \mu_i^2 \rangle$. An assumption that these are identically zero is not necessarily a good one. Moreover, when the SCF algorithm is processed with the correction of equation 2.14 in mind, the results imply that the contribution given by terms like those in equation 2.15 are not normally distributed about zero. The precise cause of such effects are unknown, but it seems to imply that the signal and noise are not as separable as they might seem.

2.3.2 Signal Degradation

With the ingenuity of the Seljak correction frustrated by unknown complications, the bias for high signal to noise data must be corrected in some other fashion. The resulting correction is far from standard in the study of spectral maps. Essentially, if the spectra cannot have all bias due to signal to noise eliminated, the next best thing is to insure that all spectra have the same amount of bias. This is accomplished by reducing the signal to noise ratio for a spectrum to be a given level. This reduction is accomplished by adding noise to the spectra in a map to reduce them all to a given threshold level. Spectra with signal to noise ratios lower than the threshold value are rejected from the SCF analysis.

In order to gauge whether the results from such a degradation are meaningful, a simulation was performed. Instead of using artificial data, a set of observational data was used.

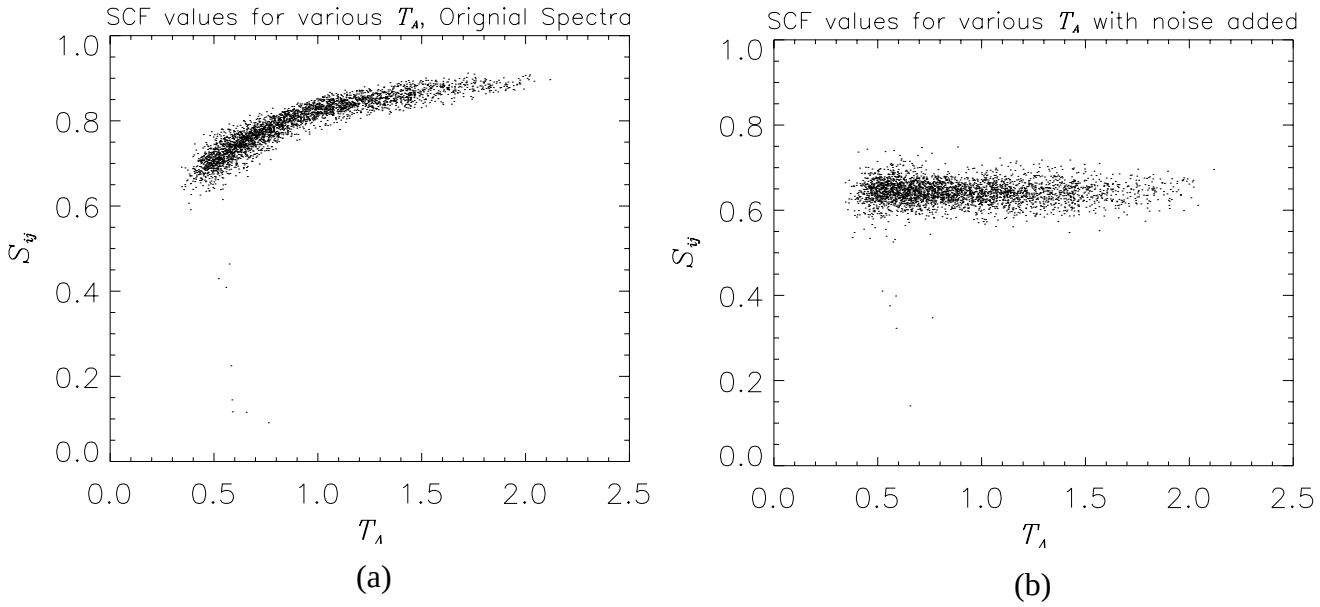

Figure 2.3: SCF as a function of antenna temperature for (a) the original data and (b) noise added to the spectra to create a uniform T/σ ratio of 5.

The degradation was performed 10 times on the same map (different sets of noise added) and the correlation function values were evaluated between the spectra. The threshold value for the signal to noise ratio was 5. The desired result was that the deviation for these values of the correlation functions would be close to zero indicating that similar results were arrived at using different sets of random noise. The results support this hope and lend credence to the method. Statistical moments of the resulting data are shown in Table 2.2. These results indicate that there is approximately a 0.2% error in using this method to eliminate the signal to noise bias.

Function	Mean	Deviation	Skewness	Kurtosis
S_{ij}	0.56	0.0021	-0.41	-0.84
S_{ij}^{ℓ}	0.57	0.0020	-0.25	-1.1
S_{ij}^{s}	0.53	0.0030	-0.46	-0.59
S_{ij}^{0}	0.54	0.0027	-0.33	-0.83

Table 2.2: Statistical moments for SCF outputs with varying noise but same T/σ

The only remaining question to ask is whether such a treatment does eliminate the bias that is observed towards correlation in higher signal to noise regions of a map. In order to explore this question, a data cube was processed with the SCF algorithm and the values of the SCF were calculated for the original data and with the noise added. Then, the results were plotted against their value of antenna temperature to determine whether any bias remained. The results appear in Figure 2.3.

As Figure 2.3 indicates, the process of adding signal to noise evens out an bias toward higher values of the antenna temperature. The mean values of the SCF are lowered a bit by this process, for some amount of correlation is lost in the process. The vertical spread in the points is not drastically altered in this process and thus the signal degradation performs the desired compensation.