Optimal Smoothing Methods for Topological Study of Galaxy Redshift Survey

Adrian L. Melott and Kurt G. Dominik
Department of Physics and Astronomy
University of Kansas

Optimal Smoothing Methods for Topological Study of Galaxy Redshift Surveys

Adrian L. Melott and **Kurt G. Dominik**
Department of Physics and Astronomy
University of Kansas
Lawrence, KS 66045

Abstract

Studying the topology of large–scale structure as a way to better understand the initial conditions which produced it has become more widespread in recent years. Studying topology in redshift space produces results compatible with the real topological characteristics of a survey (Melott, Weinberg, and Gott 1988). However, with non-periodic boundary conditions, the use of smoothing must result in the loss of information at survey boundaries. In this paper, we test different methods of smoothing samples with non–periodic boundary conditions to see which best preserves the topological features of the real distribution. We find that a smoothing method which sums only over cells inside the survey volume produces the best results among the schemes tested.

I. Introduction

The topology of large scale structures in the universe is becoming an increasingly accessible and interesting problem (for a review, see Melott 1990). With the ability to probe and survey ever deeper into the universe around us, galaxy surveys are approaching the level at which the topology of the galaxy distribution can be more reliably determined. Three distinct types of topology have been discussed as possible evolutionary results for Gaussian initial conditions. Hierarchial, or "Meatball", topology consists of isolated high density regions surrounded by voids which span the sample space, resulting in high density islands surrounded by a low density ocean. This type of topology has been modeled by Soniera and Peebles (1978). Cellular, or "bubble", topology is the inversion of Meatball topology, resulting in a high density continent spanning the survey with isolated low density lakes. This topology has been suggested for deep sky surveys (for example, see Joeveer and Einasto 1978). Finally, the third type of topology discussed is that of a Gaussian, or "Spongy", distribution. This topology can be mistaken for either of the previous two in two dimensional slices, but in fact has a unique signature. In a survey with this type of topology, the high density land masses are interconnected, and are the low density seas, resulting in both regions being able to span the survey (Gott, Melott, and Dickinson 1986). Each of these topological distributions can occur through the nonlinear evolution of a Gaussian field with different physical initial conditions (Melott, Gott, and Weinberg 1988). It is also possible that the Universe had non-Gaussian initial conditions. By using gravitational instability theory to simulate the formation of large scale structures and analyzing them topologically, we can hope to learn a great deal about the initial perturbations which created currently observable large scale structures. Our topological analysis involves an algorithm which determines the characteristic genus of selected "boundary surfaces". These surfaces separate, by density, the highest x% of the cells from the lowest (100-x)% of the cells for several values of x between zero and one, defining a *genus curve* (the genus of the boundary surface plotted against the corresponding value of x). The genus can be thought of as the number of holes in the isodensity contour, as holes in a pretzel. For a general introduction to the concepts used in this work see Melott (1990). An algorithm called Contour3D is used in conjunction with smoothing to study models and data. The theory of the genus of a Gaussian field has been discussed by Hamilton, Gott, and Weinberg (1986). They demonstrate that the form, not the amplitude, of the genus curve is the important feature determining the characteristics of the underlying distribution. Gott *et. al.* (1989) discerned the topology of real data (CfA surveys) with non-periodic boundary conditions, and found the sample to be compatible

with a spongy topology (Gaussian initial conditions with random phases). They padded the exterior of their data with a constant density equal to the mean of the sample space. To determine the error in the determined topology, they padded with a poisson exterior having an equivalent number density. Gott *et al.* (1989) studied a variety of available redshift data using mean density padding to evaluate signal and poisson padding to evaluate noise. In all cases, the data was smoothed before determining the genus curve. Smoothing is achieved by convolution with a Gaussian by Fourier transform. Usually the survey volume is embedded in a much larger volume, the large volume being filled with "padding" such as a Poisson distribution, and the entire larger volume is then treated as periodic and smoothed before the survey is used. This can introduce error at the boundary of surveys, since surveys are non-periodic. This error comes from the convolved value being dependent on neighboring density cells, roughly within L_s, the smoothing length. Thus, smoothing allows the exterior padding used in analyzing the survey to influence the survey in cells near the boundary, which distorts the characteristics of the density distribution at the boundaries. The question then arises as to how can one minimize the loss or distortion of information when smoothing samples with non–periodic boundary conditions. Could the type of padding, the amount of smoothing, or the sample size affect the topological character of real data? We attacked this question and tested several known methods of padding samples to produce periodic boundaries. We then analyzed the samples by smoothing them with the padding, and then extracting the original sample area. The original sample area is then analyzed with non–periodic boundary conditions in an effort to determine what type of padding can best minimize the loss of information at the boundaries caused by the necessity of smoothing discreet data sets.

II. Procedures

a) The Simulations

We used dynamical simulations of 128^3 particles in three dimensions evolved by an established Particle-Mesh code, evaluated by the Cloud In Cell (CIC) method on a 64^3 density mesh. Two types of simulations were specifically selected, each with different topological characteristics. The P ("Pancake") and H ("Hierarchial") simulations have power law initial conditions with the same power law index, $n = 1$. However, each has a different cutoff of low scale power in the power spectrum k, $P(k) \propto k^n$, where $P_o(k) = 0$ for $k > k_c$: $k_c = 4k_f$ for P and $k_c = 64k_f$ for H, where $k_f = \frac{2\pi}{L}$, with L being the side of the simulation square. We studied three realizations of each model. We also studied a Gaussian, $n = -1$ pure power law, "hand made", non-evolved model having no cell with $\rho < 0$. These files will be referred

to as the "samples". Thus we used "spongelike" (Gaussian), "meatball" (Hierarchial) and "bubbly" (Pancake) topologies as our samples (Melott *et al.* 1988).

b) Smoothing Techniques

To discuss how to best preserve the intrinsic topology of a sample with non–periodic boundary conditions, we performed several tests with the simulated samples described above ("bubbly" topology (P), "meatball" topology (H), and spongy topology (Gaussian)). To find the "true" or "real" topology of the samples, each was analyzed with short smoothing ($L_s = 1/64$) and periodic boundary conditions, revealing the "true" signal of the sample. For our tests, each sample was "padded" separately in four different ways to test non-periodic boundary conditions. The samples were binned on a mesh of 64^3, and were always generously padded to a mesh of 128^3. We note that there is no need to duplicate the shape of any redshift surveys, since we are measuring the intrinsic effect of boundaries. However, our results can be immediately applied to real redshift surveys.

The first method, which we termed "mean", involved padding the sample with constant density exterior to the sample equal to the mean density of the sample ($\rho = 1$ in all cases). The entire array (sample and padding) was then smoothed on a short scale ($L_s = 1/64$), and the sample was then taken back out of the padded and smoothed array. Smoothing was performed by Fourier convolution with a Gaussian

$$\rho_s\left(\vec{x}, t;\ L_s\right) = \frac{1}{2\pi}\int \rho\left(\vec{x}', t\right) \exp\left(-(\vec{x} - \vec{x}')^2/2L_s^2\right) d^2x' \ . \tag{1}$$

where L_s is the smoothing length in units of the mesh size. Ignoring non-interior vertices, the characteristic genus curve of the extracted sample was determined using the Contour3D algorithm (for a listing see Melott 1990).

The same procedure was followed for the second test, which we labelled as "Poisson", but the sample was padded instead with a Poisson exterior having the same number density of particles as the sample. The Poisson padding used was generated individually for each sample according to the number density of the sample to be padded, with particles distributed randomly, binned by CIC, and no dynamical evolution.

The third time through, called "Gaussian", the padding was a pure Gaussian (See part a)) with the same rms fluctuation amplitude as the sample. The Gaussian file used for "padding" and was created so that the density values had a pure Gaussian distribution with the same dispersion about the mean density of the samples. This means that some

cells exterior to the sample will have negative densities when using this distribution. After smoothing, any negative densities which may have been produced by the pure Gaussian exterior were set to zero.

The fourth type of padding involves surrounding the sample with zeroes for padding, then creating another 128^3 density array in which the space covered by the sample in the first array is set to a constant density of $\rho = 1$ in the second array. Both were convolved with the same Gaussian to produce smoothing. After smoothing, the array containing the original sample array is then "corrected" by dividing it by the smoothed second array, cell by cell. We labelled this as the "Ratio" test. It serves to smooth the interior of the sample without adding exterior noise through the convolution and was considered to be the most promising method from the outset.

III. Results

To choose the best smoothing method, we compared the result of subsequent analysis by the "isolated system" version of Contour3D with the result obtained when the same (periodic) simulation is smoothed periodically with the same smoothing length and analyzed with the periodic version of Contour3D.

The RMS difference between the two results is an appropriate measure. However, the result should be sensitive only to the relative error, and that error must also be compared with the intrinsic fluctuations in the model itself. Let $\sigma(\nu)$ stand for the standard deviation of $g(\nu)$ due to fluctuations from one realization to another. In reality this was evaluated at discrete values of ν which we will call σ_i. Then our measure of error

$$\epsilon = \left(\frac{1}{n-1} \sum_{i=1}^{n} \frac{(I_i - P_i)^2}{\sigma_i \, P_i} \right)^{1/2} \tag{2}$$

where I_i is the result from evaluation as an isolated system and P_i is the actual value, obtained by using periodic boundary conditions. It may be noted that this measure strongly weights fluctuations in the region where P_i is small. However, the position of the zero crossings of the genus curve is one of the best ways to discriminate various topologies, so this is an appropriate emphasis.

Each simulation was treated in two ways. In the above treatment, a smoothing length $L_s = 1$ cell was used to preserve the maximum amount of information – to have the purest possible test of the affect of boundary conditions.

However, we cannot treat observational data in this way. Previous work has shown that discreteness noise dominates the signal unless the data is smoothed with L_s equal to about the mean intergalaxy separation. We wished to model the situation in realistic samples in terms of the sampling rate and smoothing. Since the mean intergalaxy separation is about equal to the galaxy correlation length, we sampled each model by randomly selecting particles to that density. The models were then studied and compared exactly as before, except for the use of smoothing lengths $L_s = 2.28$ for H and $L_s = 3.45$ for P.

In Table I we show the results. Due to the large number of comparisons in this table we also list the Figure numbers in which the reader can see a comparison of the "true" genus curve with the one obtained by the specified smoothing procedure. The Gaussian padded by Poisson is excluded because the former was generated by direct Fourier transform and included no particles.

It can be seen that the "ratio" method is by far the most consistent performer. This is understandable in that contamination near the boundaries originates not from the padding but from nearby parts of the signal. Since mass (or galaxies) are correlated, there is a high probability that this contamination will closely resemble the signal.

We noticed a systematic effect in our best method: consistently the results show a small reduction in amplitude from the true signal. By examination of detailed slices of each, we determined that this amplitude reduction is caused by a tendency of structures near the boundaries to merge. This can be understood as due to an increase in the effective smoothing length near the boundary due to effective summation over fewer cells.

We confirmed that we could partially correct for this effect by multiplying the result by a constant: $L^3/(L - L_s)^3$. This is the ratio of total volume of the box to the volume not within L_s of a boundary. This improved the fit as shown in the scaled results in the last two rows of the table. (Of course, for differently shaped volumes, the geometrical factor would be different.)

It might be argued that these results are inapplicable to real redshift surveys, which have differently shaped volumes and variable sampling rates in magnitude–limited samples. However, our methods are not dependent on the shape of the survey volume. All finite surveys have boundaries, and there is no evidence that periodic boundary conditions are applicable to the universe. For magnitude–limited surveys, the ratio of surface to volume will be large; the effects we examine here will be even more important.

IV. Conclusions

We have conducted a systematic study of various methods of smoothing finite samples to recover the topology (Gott *et al.* 1989). We conclude that the so–called "ratio" method, described below, produces the most nearly correct results:

The survey should be embedded in a cube so that no part is close to a face (a few smoothing lengths). Areas outside the survey are assigned value zero. The resulting array is smoothed by Fourier convolution with a Gaussian of the appropriate scaling (usually the mean intergalaxy separation).

Another cube is prepared with entries 1 inside the survey volume and 0 outside, and smoothed with the same smoothing length. The result of the first smoothed cube (above) is divided cell-by–cell by the entries in the second cube.

After analysis of the survey by Contour3D (interior boundary conditions), the amplitude of the genus curve is multiplied by the ratio of the survey volume to that part of the volume not within one smoothing length of the boundary.

This method produces a more reliable result than any used to date, and can be recommended for future work. We hope that within a decade surveys will be available which are so large that one can afford to ignore regions near the boundary.

V. Acknowledgements

The authors are grateful for useful discussions with David Weinberg. We knowledge support under NASA grant NAGW–1288, NSF grant AST–9021414 (including a Research Experiences for Undergraduates supplement), and the University of Kansas GRF fund. K.D. would like to thank the University of Kansas Honors Program for a Summer 1991 Research Grant.

REFERENCES

Gott, J. R., Melott, A. L., and Dickinson, M. 1986, Ap.J., 306, 341

Gott, J. R., Miller, J., Thuan, T. X., Schneider, S. E., Weinberg, D. H., Gammie, C., Polk, K., Vogeley, M., Jeffery, S., Bhavsar, S. P., Melott, A. L., Giovanelli, R., Haynes, M. P., Tully, R. B., and Hamilton, A. J. S. 1989, Ap.J., 340, 624

Joeveer, M., and Einasto, U. 1978, in *IAU Symposium 79: The Large Scale Structure of the Universe*, ed. M. S. Langair and J. Einasto (Dordrecht: Reidel), 241

Melott, A. L. 1990, Phys. Rep., 193(1)

Melott, A. L., and Shandarin, S. F. 1989, Ap.J, 343, 26

Melott, A. L., Weinberg, D. H., and Gott, J. R. 1988, Ap.J., 328, 50

Soniera, R. M., and Peebles, P. J. E. 1978, Astron.J., 83, 845

Table 1

A summary of the values of ϵ as defined in equation (2) and the Figure numbers which display the results.

	Padding Method			
Model Type	Mean	Poisson	Gaussian	Ratio
Gaussian	0.68 (Fig. 1)	–	0.31 (Fig. 2)	0.22 (Fig. 3)
Pancake (P)	10.5 (Fig. 4)	6.9 (Fig. 5)	54. (Fig. 6)	0.50 (Fig. 7)
Hierarchial (H)	3.4 (Fig. 8)	3.4 (Fig. 9)	12.3 (Fig. 10)	0.48 (Fig. 11)
Pancake Sparse	9.4 (Fig. 12)	9.4 (Fig. 13)	0.41 (Fig. 14)	0.50 (Fig. 15)
Hierarchial Sparse	2.2 (Fig. 16)	2.4 (Fig. 17)	0.17 (Fig. 18)	0.43 (Fig. 19)
Pancake Sparse Scaled	–	–	–	0.41 (Fig. 20)
Hierarchial Sparse Scaled	–	–	–	0.18 (Fig. 21)

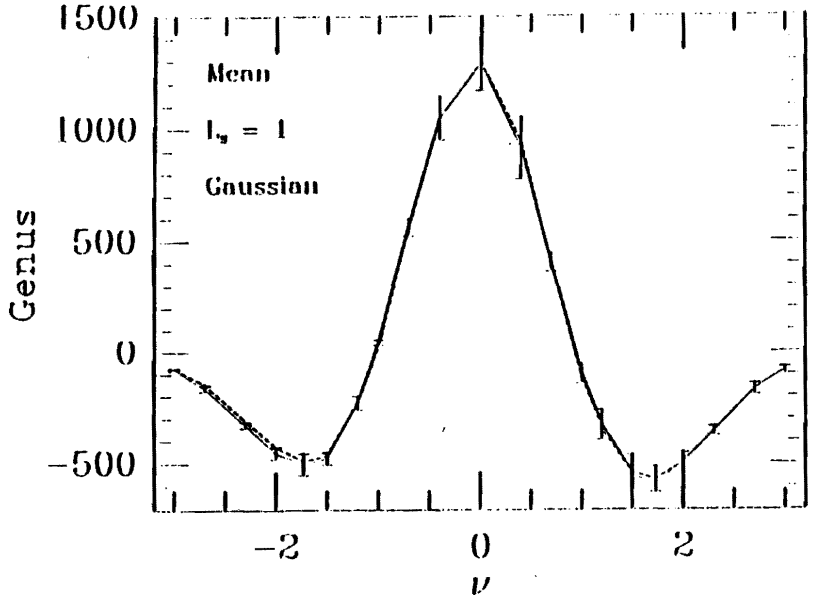

Figure 1

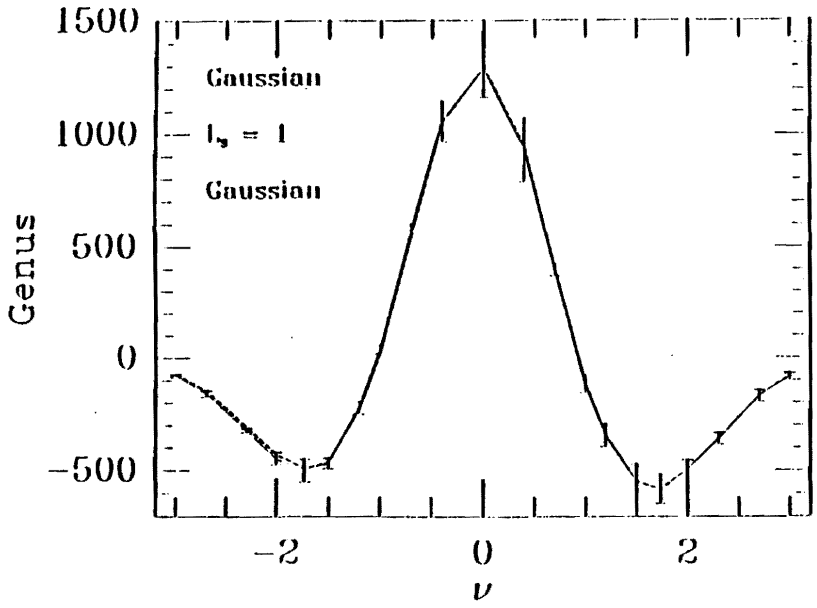

Figure 2

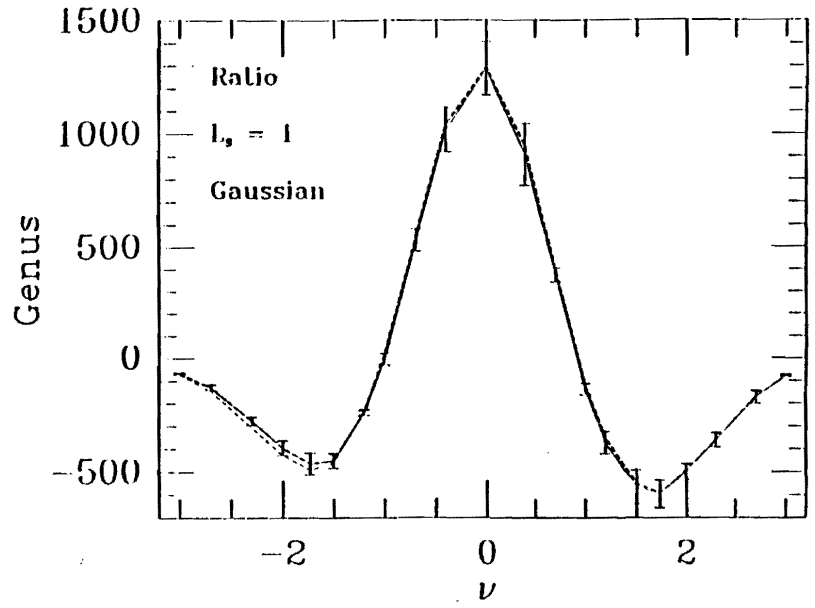

Figure 3

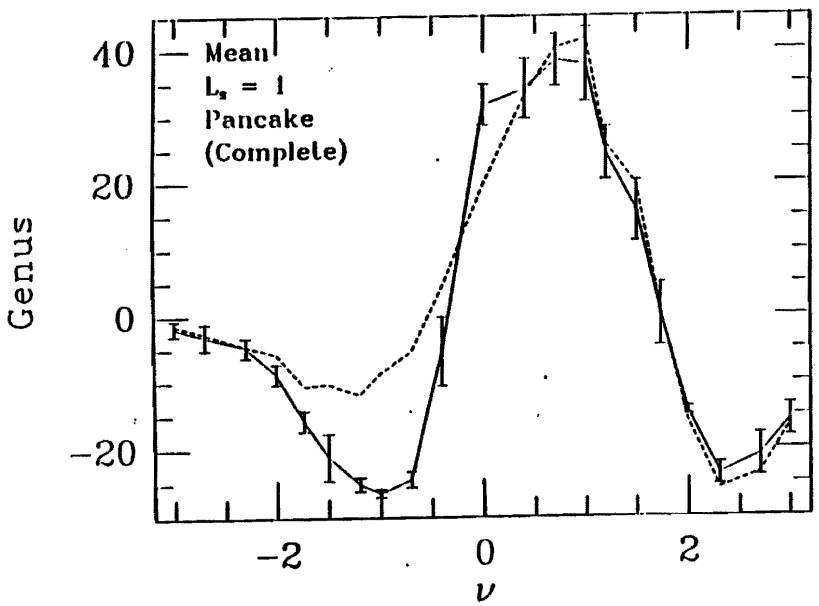

Figure 4

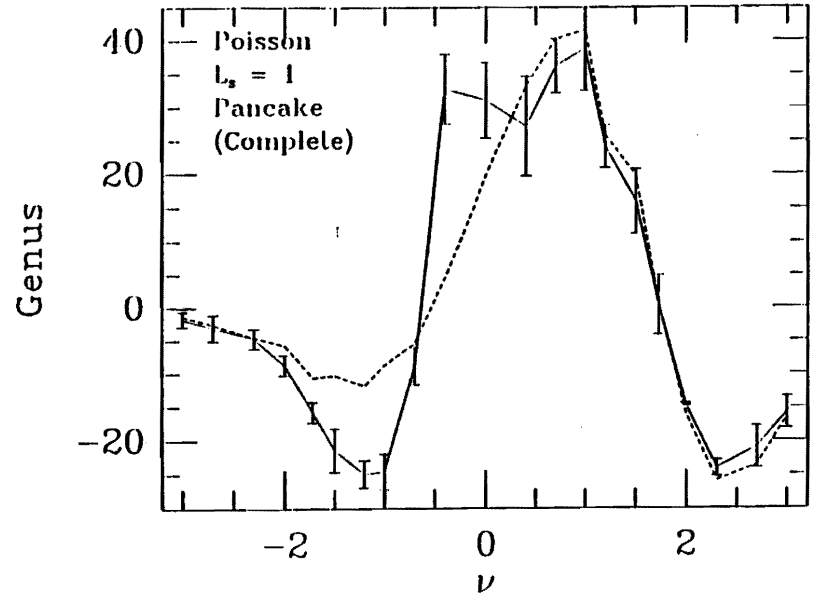

Figure 5

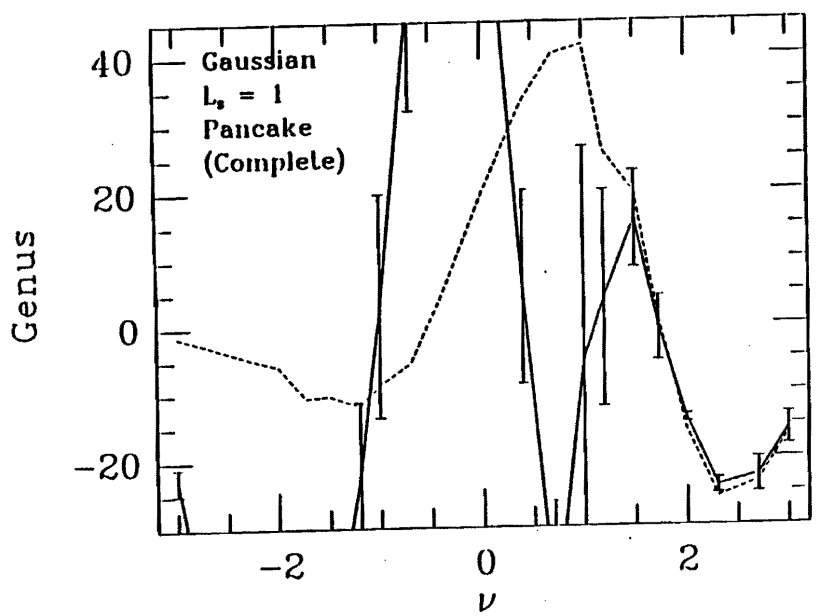

Figure 6

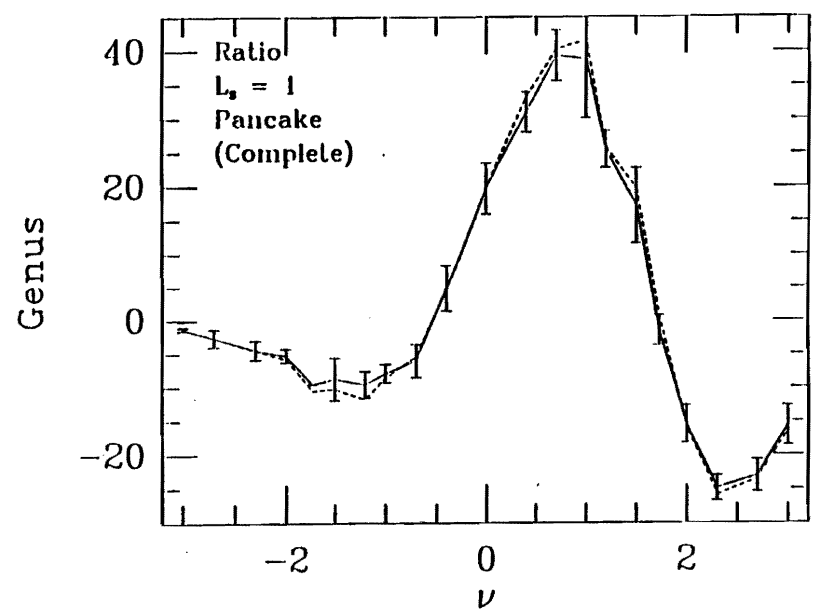

Figure 7

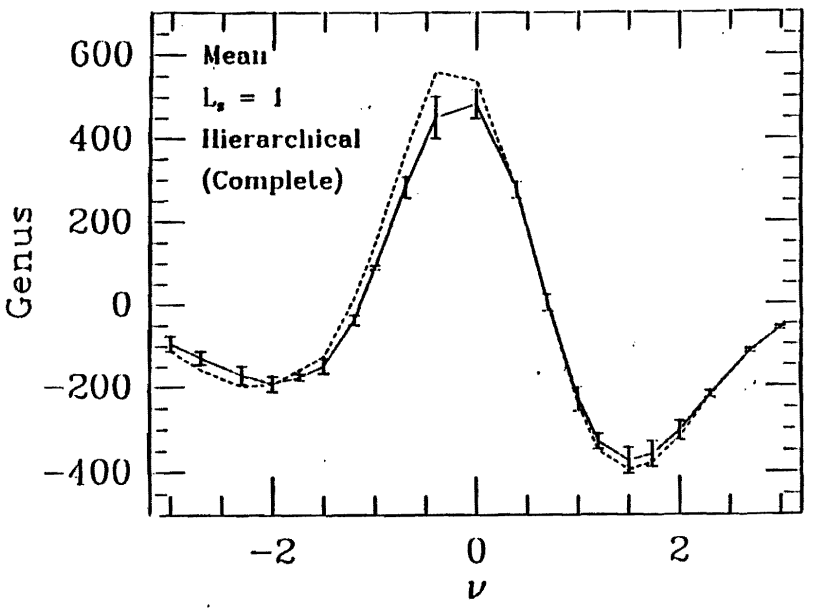

Figure 8

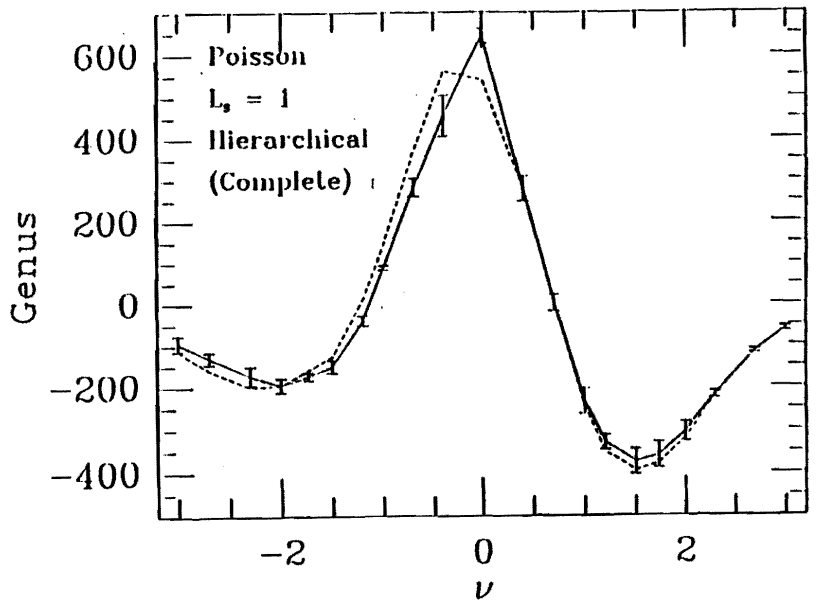

Figure 9

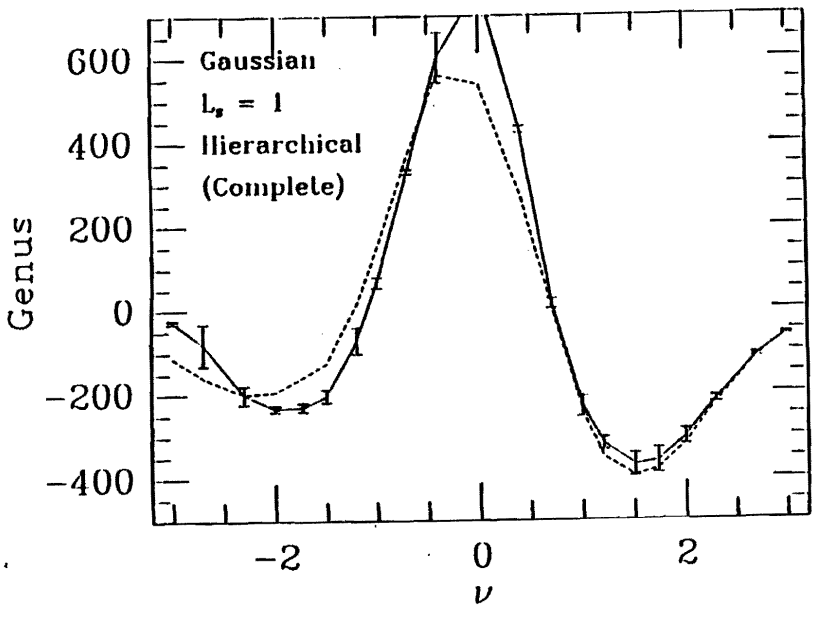

Figure 10

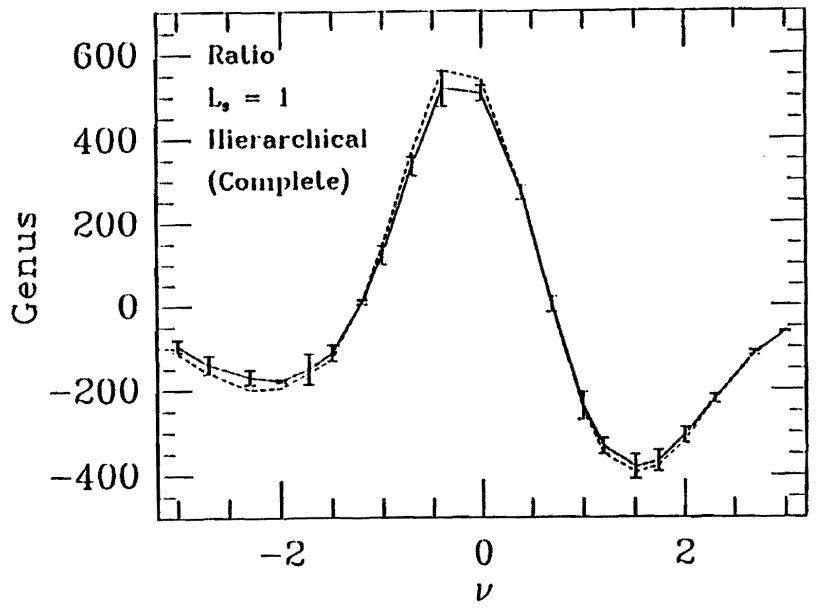

Figure 11

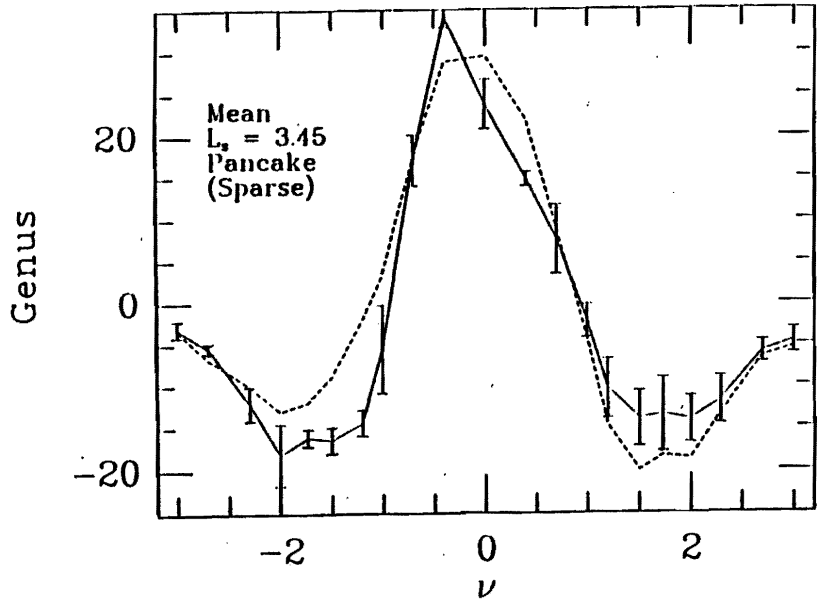

Figure 12

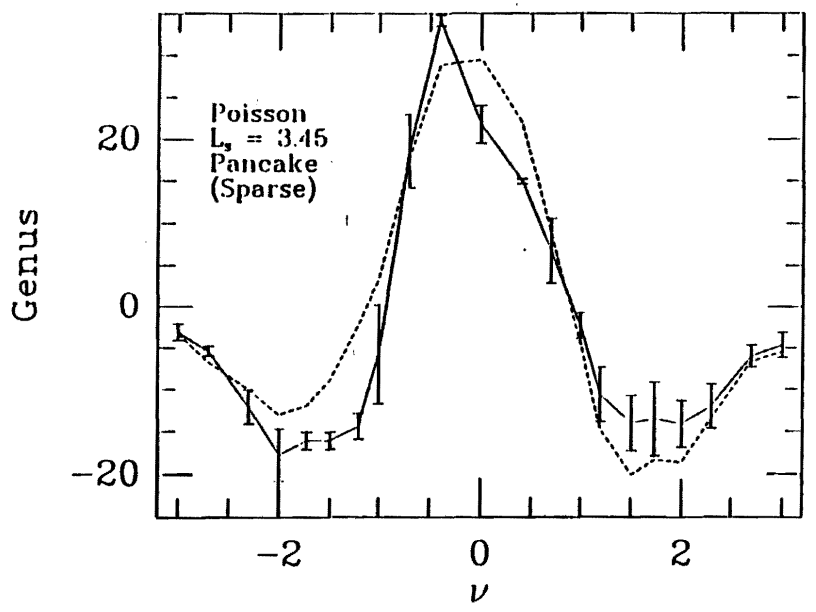

Figure 13

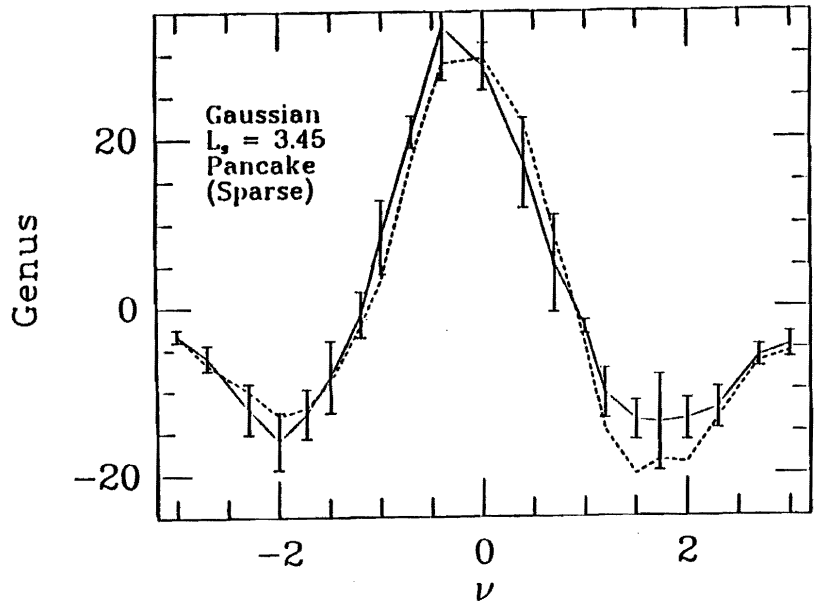

Figure 14

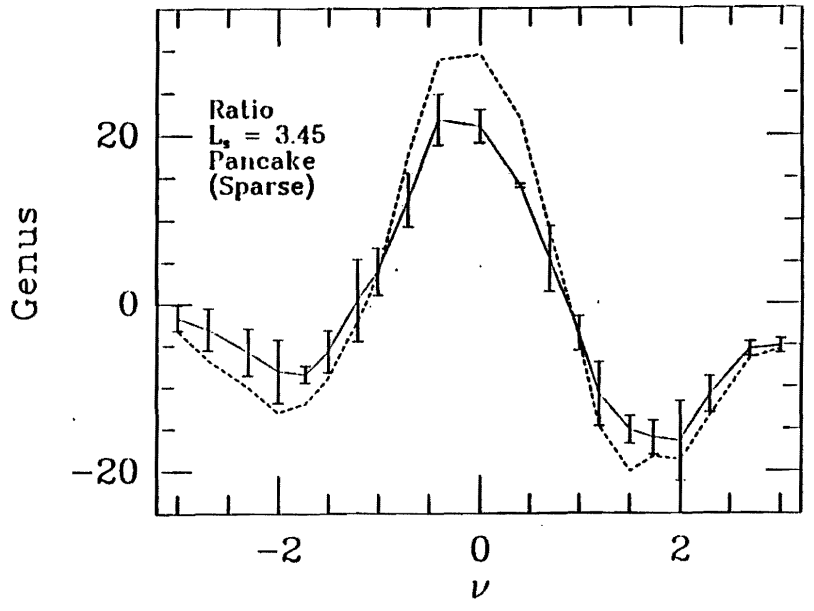

Figure 15

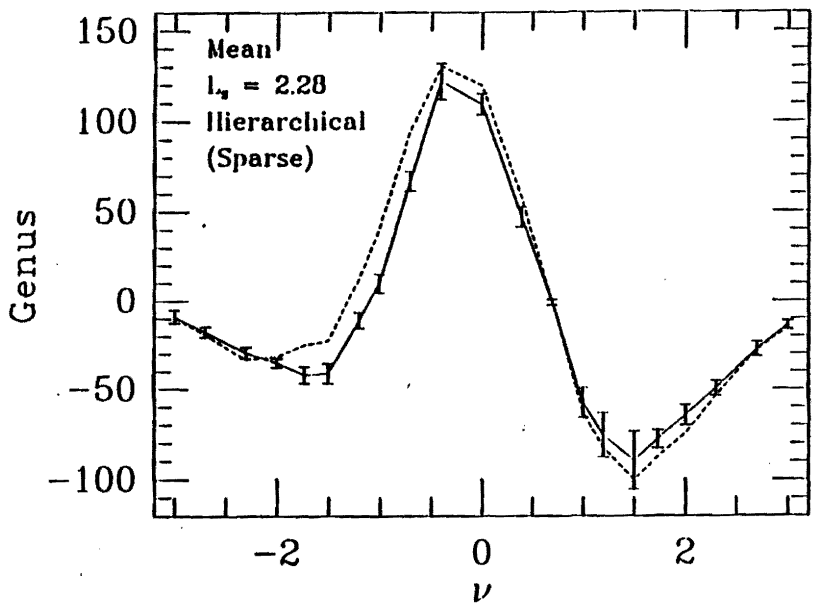

Figure 16

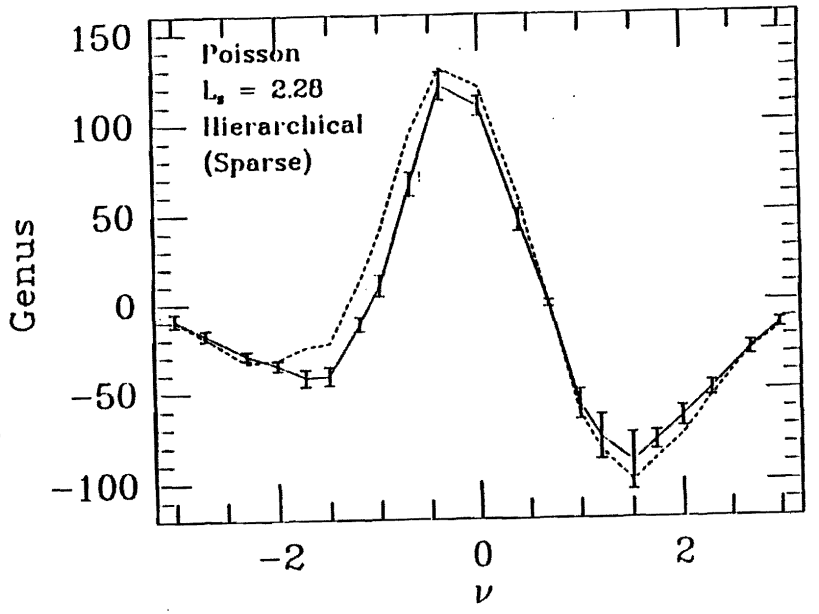

Figure 17

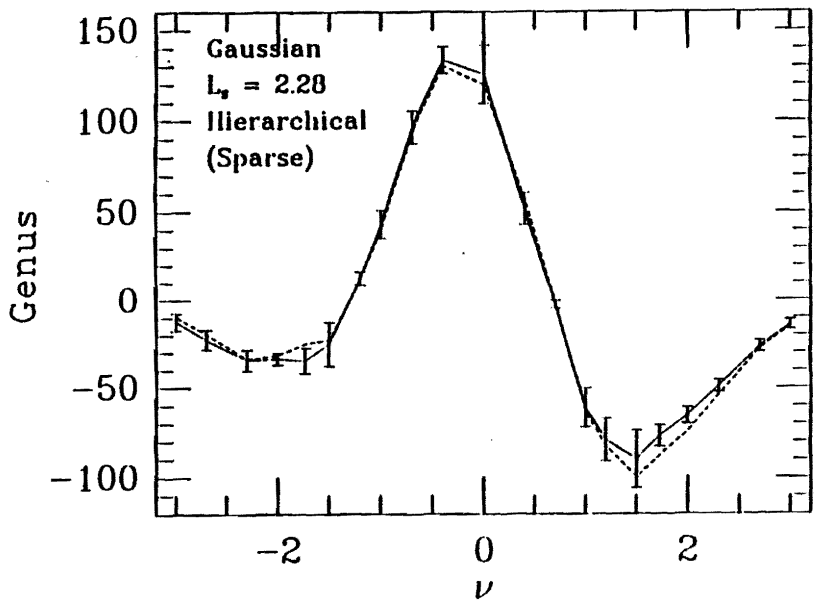

Figure 18

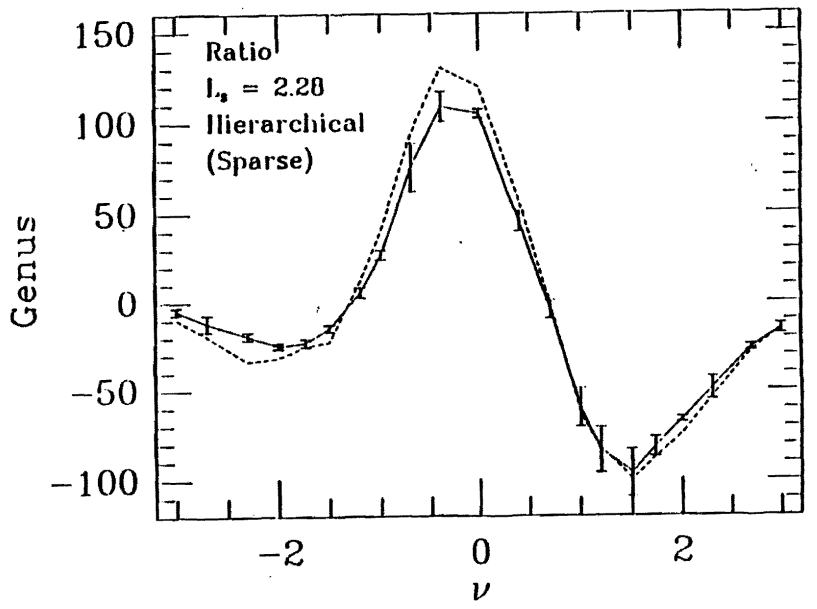

Figure 19

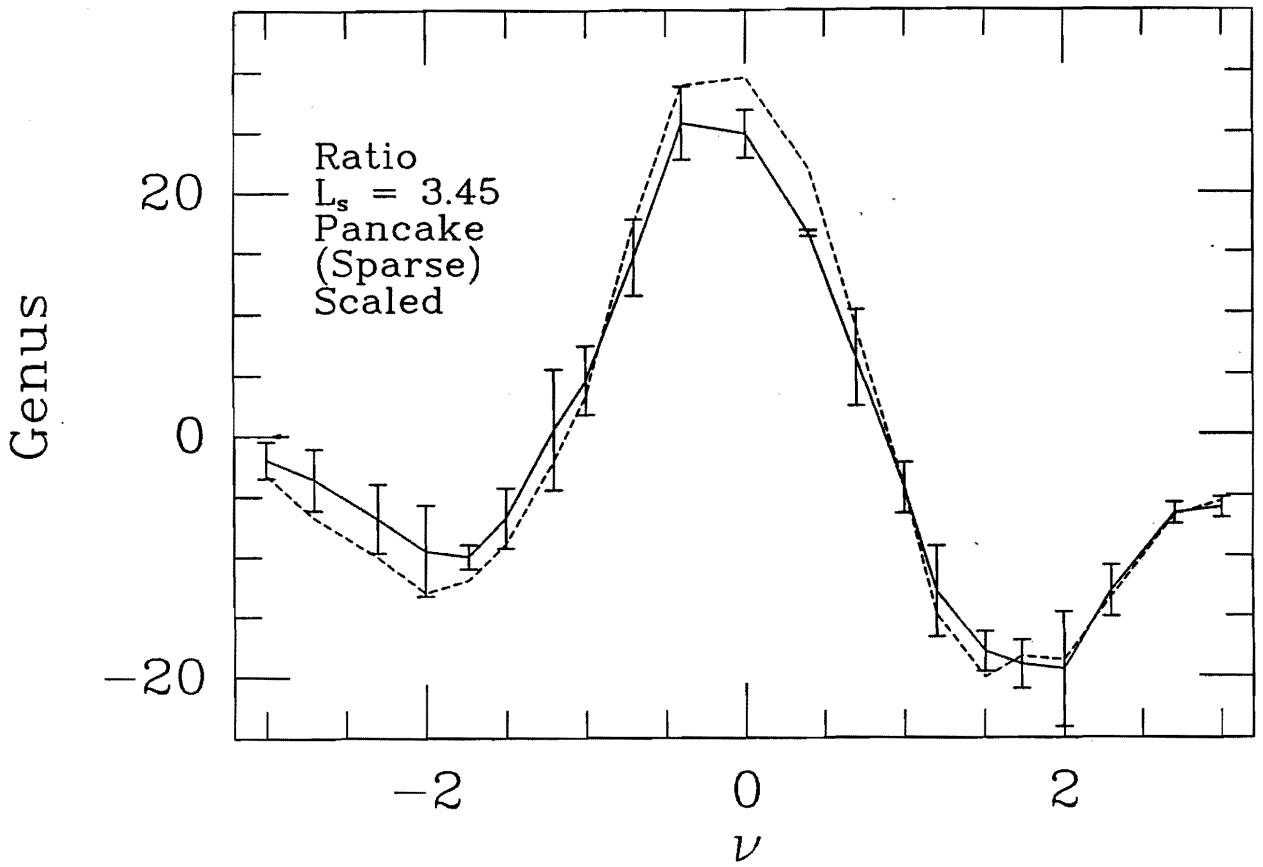

Figure 20

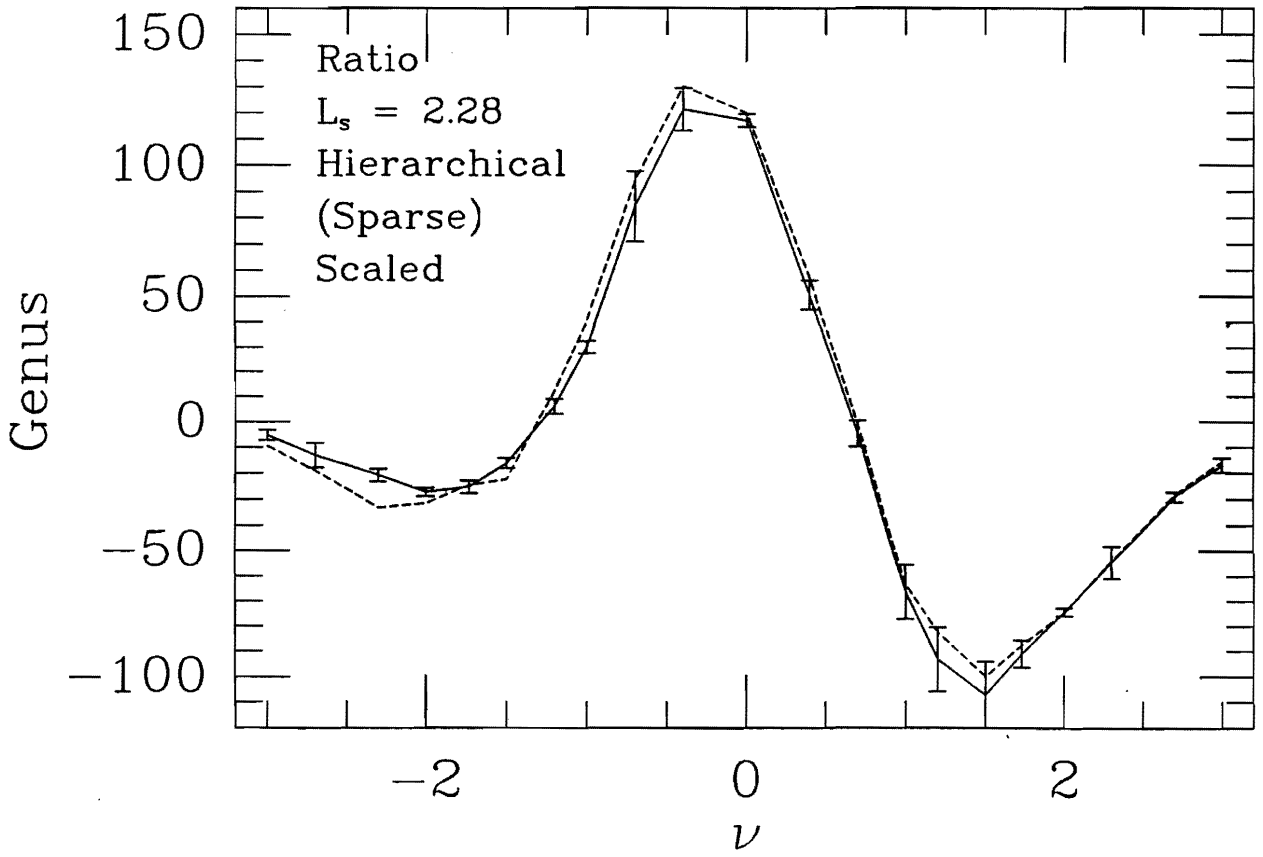

Figure 21