Ways to Improve Your Correlation Functions

A. J. S. Hamilton
Joint Institute for Laboratory Astrophysics
and Department of Astrophysical, Planetary and Atmospheric Sciences
University of Colorado, Boulder

ABSTRACT. This paper describes a number of ways to improve on the standard method for measuring the two-point correlation function of large scale structure in the Universe. Issues addressed are: (1) The problem of the mean density, and how to solve it; (2) How to estimate the uncertainty in a measured correlation function; (3) Minimum variance pair weighting; (4) Unbiased estimation of the selection function when magnitudes are discrete; (5) Analytic computation of angular integrals in background pair counts.

1. The Mean Density Problem

It is widely thought that the accuracy of the correlation function ξ is fundamentally limited by uncertainty in the mean density. Actual, this notion is *false* (§1.2), although it is true for the commonly used estimator of ξ (§1.1).

1.1 The Problem

The statistic commonly used to estimate ξ from a catalog of galaxies is (e.g. Davis and Peebles 1983)

$$\xi_{\rm est} = \frac{<NN><W>}{<NW><N>} - 1 \qquad (1)$$

where N represents real galaxies, W is the catalog window, and $<>$ denotes averaging over all points in the catalog; for $<NN>$ and $<NW>$ the averaging is over all pairs lying in an interval of separations r. The observed galaxy density N is the true galaxy density n times the catalog window W (strictly, observed discrete galaxies are taken to be a Poisson process superimposed on this). In terms of the true galaxy overdensity $\delta \equiv (n - \bar{n})/\bar{n}$, where $\bar{n}$ is the true mean galaxy density of the Universe, the observed galaxy density N is

$$N = \bar{n} W (1 + \delta) \quad . \qquad (2)$$

To see how good the estimate (1) of ξ is in a realistically unfair sample, introduce the following notations (3)-(5). Let $\bar{\delta}$ be the mean overdensity in the catalog

$$\bar{\delta} \equiv \frac{<W\delta>}{<W>} \qquad (3)$$

and let ψ denote the galaxy-catalog correlation function

$$\psi_{12} \equiv \frac{<W_1 \delta_1 W_2>}{<W_1 W_2>} \quad . \tag{4}$$

In a fair sample, the mean overdensity $\bar{\delta}$ and the galaxy-catalog correlation function ψ would be identically zero, but they are not necessarily zero in reality. Let $\hat{\xi}$ denote the windowed galaxy-galaxy correlation function

$$\hat{\xi}_{12} \equiv \frac{<W_1 \delta_1 W_2 \delta_2>}{<W_1 W_2>} \quad . \tag{5}$$

While $\hat{\xi}$ is not necessarily equal to the true correlation function ξ of the Universe, it is at least the 'true' correlation function of the sample, which is presumably the next best thing.

In terms of $\bar{\delta}$, ψ, and $\hat{\xi}$, the standard estimate (1) of the correlation function ξ is

$$\xi_{est}(r) = \frac{\hat{\xi}(r) + \psi(r) - \bar{\delta} - \psi(r)\bar{\delta}}{(1 + \psi(r))(1 + \bar{\delta})} \quad . \tag{6}$$

The problem with equation (6) is that it contains not-necessarily-vanishing terms $\psi(r) - \bar{\delta}$ which are of first order in overdensity δ, whereas the thing you want, the sample correlation function $\hat{\xi}$, is of second order in δ. This is a severe drawback of the estimator (1) for ξ in the linear regime of small δ.

1.2 A Solution, Part 1

A better estimate of ξ is

$$\xi_{est} = \frac{<NN><WW>}{<NW>^2} - 1 \tag{7}$$

in which the brackets $<>$ in both numerator and denominator denote averaging over pairs in an interval of separations r. In terms of the galaxy-catalog correlation function ψ and the sample galaxy-galaxy correlation function $\hat{\xi}$ defined by equations (4) & (5), the estimate (7) is

$$\xi_{est}(r) = \frac{\hat{\xi}(r) - \psi(r)^2}{(1 + \psi(r))^2} \tag{8}$$

which differs from the sample correlation function $\hat{\xi}$ only by terms which are of second order in overdensity δ.

The advantages of the estimator (7) over the standard estimator (1) for ξ are:

(a) Accuracy, especially in the large scale, linear regime;

(b) Reliability in the presence of unfairness, especially with not-unbiased (e.g. minimum variance, §3) pair weightings;

(c) Peace of mind: there is no need to measure the mean density $<N>/<W>$ as a separate operation; equation (7) specifies that the 'correct' mean density to use in place of the $<N>/<W>$ in equation (1) is $<NW>^2/<WW>$, a quantity which it is to be noted varies with separation r.

1.3 SOLUTION, PART 2

The ψ^2 term in equation (8) represents large scale variance which is inevitably missing in a finite catalog; its presence is symptomatic of the familiar problem that using the sample mean leads to an underestimate of the sample variance. Although the galaxy-catalog correlation ψ should be zero in the mean over many samples, its variance $<\psi^2>$ should be positive in the mean. Physically, $<\psi^2>$ represents the mean fractional excess of galaxies clustered around a galaxy on the scale of the catalog. If one imagines evaluating the correlation function by sitting on galaxies and counting neighbors, the mean density at infinity should be determined not from the total number of galaxies in the catalog volume, but rather from the number of galaxies less the mean excess of galaxies clustered around a galaxy.

One way to correct for the missing variance is to use an alternate estimator

$$\xi_{\rm est} = \frac{<NN><WW>}{<NW>^2 - <\Delta(NW)^2>} - 1 \qquad (9)$$

which in terms of $\hat{\xi}$ and ψ and its variance $<\psi^2>$ is

$$\xi_{\rm est}(r) = \frac{\hat{\xi}(r) - \psi(r)^2 + <\psi(r)^2>}{(1+\psi(r))^2 - <\psi(r)^2>} \quad . \qquad (10)$$

The variance $<\Delta(NW)^2>$ in equation (9) may be computed from the fluctuations in $<NW>$, using methods similar to those described in §2. Another way to correct for missing large scale variance is given by Hamilton (1993).

2. Estimating Errors in the Correlation Function

2.1 MATHEMATICS

The window W can be imagined as a set of weights W_i attached to every tiny volume element of the Universe. For an observed subsample, the weights W_i are nonzero only over the observed region. For the entire population, the Universe, the weights $W_{i,\rm pop}$ are finite everywhere, but infinitesimal compared to the sample weights W_i.

The correlation function $\xi(W_i)$ measured in a sample differs from the true correlation function $\xi(W_{i,\rm pop})$ by an error $\Delta\xi$

$$\Delta\xi = \xi(W_i) - \xi(W_{i,\rm pop}) \quad . \qquad (11)$$

Expanding the error $\Delta\xi$ as a Taylor series to second order in the weights gives

$$\Delta\xi = \sum_i (W_i - W_{i,\rm pop}) \frac{\partial\xi}{\partial W_i}\bigg|_{\rm pop} + \frac{1}{2}\sum_{ij}(W_i - W_{i,\rm pop})(W_j - W_{j,\rm pop}) \frac{\partial^2\xi}{\partial W_i \partial W_j}\bigg|_{\rm pop} \quad . (12)$$

Using the facts that (a) $\xi(W_i)$ is unchanged by rescaling $W_i \to \lambda W_i$, (b) ξ is a quadratic function of the weights W_i (at least for the estimator [7]), and (c) $W_{i,\rm pop}$ is infinitesimal compared to W_i, eliminates most of the terms in equation (12), reducing it to

$$\Delta\xi = \frac{1}{2}\sum_{ij} W_i W_j \frac{\partial^2\xi}{\partial W_i \partial W_j}\bigg|_{\rm pop} \qquad (13)$$

which then reduces further to

$$\Delta\xi = \sum_{\text{distinct } ij} W_{ij} \frac{\partial \xi}{\partial W_{ij}}\bigg|_{\text{pop}} \qquad (14)$$

where $W_{ij} \equiv W_i W_j$. Expression (14) makes clear the fact that the error $\Delta\xi$ is truly a derivative with respect to *pairs*. Approximating the population derivative of ξ in (14) by the sample derivative, and again using the fact that $\xi(W_i)$ is quadratic in W_i, permits equation (14) to be rearranged as a sum over volume elements i rather than pairs ij:

$$\Delta\xi = \sum_i \Delta\xi_i \qquad \text{with} \qquad \Delta\xi_i = \frac{1}{2} W_i \frac{\partial \xi}{\partial W_i} \quad . \qquad (15)$$

Note the important factor of $^1/_2$ in equation (15), which in effect causes pairs to be counted once, not twice. The variance of ξ is then

$$<\Delta\xi^2> = \sum_{ij} \Delta\xi_i \Delta\xi_j \quad . \qquad (16)$$

Characteristically, the variance (16) increases as pairs ij of greater and greater separation are included, reaches a maximum, then declines to exactly zero when all pairs are included. The declining to zero is a consequence of approximating the population derivatives of ξ with the sample derivatives of ξ. To solve the problem, only pairs ij separated by some finite distance should be included.

For the estimator (7), the contribution $\Delta\xi_i$ to the error in ξ from volume element i is

$$\Delta\xi_i = (1 + \xi_{\text{est}}) \left[\frac{<N_i N>}{<NN>} - \frac{<N_i W>}{<NW>} - \frac{<NW_i>}{<NW>} + \frac{<W_i W>}{<WW>} \right] \quad . \qquad (17)$$

2.2 Suggested Step-by-Step Error Analysis

(a) Divide the catalog into many subregions i.

(b) Estimate ξ from equation (7), and compute $\Delta\xi_i$ for each subregion from equation (17).

(c) Compute the variance $<\Delta\xi^2>$ from equation (16), including pairs ij of subregions of greater and greater separation, until the variance reaches a maximum.

(d) The 1-sigma error in ξ is the square root of this variance.

3. Minimum Variance Pair Weighting

3.1 Mathematics

An unbiased estimate of the correlation function ξ is gotten in principle by weighting each point inversely with the selection function Φ at the point, so that all volume elements count equally. Unfortunately this leads to a noisy estimate of ξ from regions where the selection function Φ is small. To the extent that the selection function is uncorrelated with the true galaxy density, the most accurate estimate of ξ is obtained by reweighting the unbiased weighting of pairs inversely with the variance $<\Delta\xi^2>$ of ξ, so that a (real or background) pair 12 is weighted

$$w_{12} = \frac{1}{\Phi_1 \Phi_2 <\Delta\xi^2>} \quad . \qquad (18)$$

If the only source of uncertainty in ξ comes from the fact that the sample is a finite subset of the Universe, then the expected covariance between ξ's at separations r_{12} and r_{34} is

$$\begin{aligned} <\Delta\xi_{12}\,\Delta\xi_{34}> &= <\delta_1\delta_2\delta_3\delta_4> - <\delta_1\delta_2><\delta_3\delta_4> \\ &\propto \int (\xi_{13}\xi_{24} + \xi_{14}\xi_{23} + \eta_{1234})\, dV_3 \end{aligned} \tag{19}$$

the integral being carried over all possible separations of point 3 from point 1. An important point to notice is that the ξ's in the integrand of (19) have delta functions at zero separation, because of the discreteness of galaxies. These delta functions cause equation (19) to take the general form

$$<\Delta\xi^2> \propto \Phi^{-2} + 2\Phi^{-1}J + K \tag{20}$$

in a region where the selection function is Φ. The Φ^{-2} term in equation (20) comes the case where pairs 12 and 34 coincide, the Φ^{-1} term from cases where 12 and 34 share a point in common, and the constant term from cases where 12 and 34 are disjoint pairs. Equation (20) yields the pair weighting

$$w_{12} = \frac{1}{1 + 2\Phi J + K\Phi^2} \quad . \tag{21}$$

Generally one is interested not in ξ at some precise separation, but rather averaged over some range of separations; or one might be interested in the power spectrum, or the harmonics of ξ, or such like. In that case equation (19) should be integrated with the desired kernel functions over the desired ranges of separations. The result is again equations of the form (20) and (21), but with different values of the coefficients J and K. The bad news is that a calculation of J and K from equation (19) is generally tricky and uncertain, and in any case suspect because an observationally measured ξ may be subject to other sources of uncertainty ignored in equation (19). The good news is, the calculation is unnecessary. A practical solution is to proceed empirically, using the form (21) as a template for an approximate pair weighting, the free parameters of which would be determined empirically by minimizing the observed variance $<\Delta\xi^2>$ computed for example using the method of §2. This is the approach suggested in §3.2 below.

3.2 Suggested Near Minimum Variance Weighting

A simple approximation to the minimum variance pair weighting which should not be too bad in practice would be to weight every (real and background) pair 12 by $w_{12}=w_1w_2$, the weight w_i at each point i where the selection function is Φ_i being

$$w_i = \frac{1}{1 + \Phi_i J} \quad . \tag{22}$$

The quantity J in equation (22) is likely to be different for different pair separations r. Consideration of the behavior of the integral (19) suggests that a reasonable guess would be to take

$$J = Cr^c \tag{23}$$

with $c \approx 3-\gamma$ if $\xi \propto r^{-\gamma}$. The free parameters C and c in (23) would be determined empirically by varying them until the computed variance $<\Delta\xi^2>$ of ξ, or of whatever integral over ξ is the quantity of interest, is minimized. Clearly the approximation (23), and perhaps also (22), could be refined if deemed necessary.

4. The Selection Function and the Discreteness of Magnitudes

4.1 PROBLEM

Turner's (1979) classic inhomogeneity-insensitive method of measuring the selection function is found empirically to be sensitive to bin size.

4.2 DIAGNOSIS

Strauss, Yahil & Davis (1991) correctly attribute the sensitivity of Turner's method to bin size to the fact that magnitudes given in catalogs are discretely, not continuously, distributed.

The periodicity of listed magnitudes (e.g. Zwicky gives magnitudes mostly to 0.1) translates into a periodic ripple in the derived selection function (with period 0.1 magnitudes in Zwicky's case).

4.3 SOLUTION

The fix is simple: just choose a bin size equal to the period (0.1 magnitudes in Zwicky's case), so the periodic ripple has no effect. Note that the worst possible binning is exactly half a period, since successive samplings of the selection function are then maximally out of phase.

5. Better Backgrounds

Oftentimes a catalog window is the product of a radial selection function and an angular window which is one inside, zero outside, a boundary composed of a set of arc segments. Angular integrals in such a case can be done *analytically*, and backgrounds can then be integrated quickly and accurately.

[At the conference, a binder was exhibited containing listings of Fortran code to compute angular integrals analytically. The relevant mathematics is given by Hamilton (1993, Appendix).]

References

Davis, M., & Peebles, P. J. E. 1983, ApJ 267, 465.
Hamilton, A. J. S. 1993, ApJ, submitted.
Strauss, M. A., Yahil, A., & Davis, M. 1991, PASP, 103, 1012.
Turner, E. L. 1979, ApJ, 231, 645.