Mon. Not. R. astr. Soc. (1986) **223,** *Short Communication,* 21P–27P

Uncertainties in the cluster–cluster correlation function

E. Nigel Ling[1], C. S. Frenk[2] and John D. Barrow[1]

[1] *Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH*
[2] *Department of Physics, The University, South Road, Durham DH1 3LE*

Accepted 1986 September 24. Received 1986 September 5; in original form 1986 July 22

Summary. We apply the bootstrap resampling technique to estimate sampling errors and significance levels of the two-point correlation functions determined for a subset of the CfA redshift survey of galaxies and a redshift sample of 104 Abell clusters. We also calculate the angular correlation function for a sample of 1664 Abell clusters. We find the standard errors in $\xi(r)$ for the Abell data to be considerably larger than quoted 'Poisson errors'. Our best estimate for the ratio of the correlation length of Abell clusters (richness class $R \geqslant 1$, distance class $D \leqslant 4$) to that of CfA galaxies is $4.2^{+1.4}_{-1.0}$ (68 percentile error). The enhancement of cluster clustering over galaxy clustering is statistically significant in the presence of resampling errors. The uncertainties found do not include the effects of possible systematic biases in the galaxy and cluster catalogues and could be regarded as lower bounds on the true uncertainty range.

1 Introduction

The study of the Universe's large-scale structure has involved the application of various statistical measures of galaxy clustering. Of these, the two-point correlation function, $\xi(r)$, of galaxies separated by distance r has proved the most popular because of the ease with which it can be computed and the simple scaling relations that exist between its two- and three-dimensional forms. It has been found that galaxies are strongly clustered on scales less than $10\,h^{-1}$ Mpc (the Hubble constant is $H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$) with a power-law correlation function

$$\xi_g(r) = (r/r_0)^{\gamma}, \tag{1}$$

where $\gamma \simeq -1.8$ and the correlation length is $r_0 \simeq 5\,h^{-1}$ Mpc (Davis & Peebles 1983). Recent computations of the two-point correlations of rich clusters, $\xi_c(r)$, have revealed them to possess the functional form (1), also with $\gamma \simeq -1.8$ but with a significantly larger amplitude $r_0 \simeq 25\,h^{-1}$ Mpc (Bahcall & Soneira 1983; Klypin & Kopylov 1983; Postman, Geller & Huchra 1986).

Kaiser (1984) suggested that the enhancement of cluster–cluster correlations could be explained if the comparatively rare rich clusters formed only at places where the density exceeds a high threshold (Peacock & Heavens 1985; Bardeen *et al.* 1986; Coles 1986). If these high-contrast regions are identified with rich clusters then, so long as the underlying density fluctuations

constituted a Gaussian random field initially, the cluster–cluster two-point correlations are amplified with respect to the two-point correlations of the total density field, with $\xi_c \propto \xi_g$. This amplification is seen in numerical simulations of scale-free hierarchical models of galaxy formation and clustering, as well as in simulations of neutrino-dominated universes; in the latter case, the predicted enhancement is much too small to account for what is observed (Barnes *et al.* 1985). Attempts have also been made to explain cluster correlations in the context of vacuum string theories which naturally contain non-Gaussian fluctuations (Turok 1985).

A current model for the origin of cosmic structure assumes that galaxies form only near high peaks of the smoothed linear density field, where the latter is provided by a background of cold dark matter in an Einstein–de Sitter universe (Blumenthal *et al.* 1984; Davis *et al.* 1985). In this model, neither the galaxies nor the clusters are fair tracers of the mass and the ratio of their correlation amplitudes reflect their relative bias with respect to the mass distribution. In a recent investigation, White *et al.* (1986) showed that while the predicted correlation length of rich clusters in this model is five times that of the dark matter distribution, it is only 2.1 times that of the galaxy distribution.

It is therefore now of considerable importance to quantify as precisely as possible the significance of the observational data, particularly that of the ratio of the clustering strengths of galaxy and cluster clustering. This is not straightforward. The general form for the uncertainty in the estimate of $\xi(r)$ is complicated and involves two-, three- and four-point correlation functions; simplified expressions have been derived, but only for the limiting situation where $\xi \ll 1$ (Peebles 1973; Kaiser 1986). The 'Poisson error bars' often quoted assume an uncorrelated distribution and thus may seriously underestimate the true uncertainties. In practice, a further complication arises from the need to estimate the mean density of the sample from the same data set used to estimate ξ. In the case of the Abell cluster sample, in addition to statistical uncertainties, there may well be systematic biases associated with the subjective nature of the selection procedure, misidentifications, obscuration and projection effects, and the use of estimated rather than measured redshifts (Lucey 1981, 1983; Postman *et al.* 1986). Such biases may be quite complex and will not be fully understood until current photometric and spectroscopic surveys are completed.

In this paper we are concerned exclusively with sampling uncertainties; we show how a statistical significance can be associated with the computations of the cluster–cluster correlation function using modern resampling techniques. Ideally, to estimate sampling uncertainties, we would like to have a large number of independent samples. In the case of non-repeatable observations like those of rich cluster clustering there are well-documented techniques for generating pseudo data sets from the original sample. The magnitude of the variance of any quantity determined over the ensemble of pseudo data sets allows us to assess the robustness of the original data set to any source of sampling error*.

If this variance is found to be large then it indicates a lack of robustness in the original data set with respect to the statistical parameter being evaluated. An application of these techniques to calculations of the angular two-point correlation function of the Zwicky catalogue by Barrow, Bhavsar & Sonoda (1984) found standard errors that significantly exceed the quoted measurement errors. Resampling methods also offer a good means of associating a statistical significance to measures of filamentary or cellular pattern (Barrow & Bhavsar 1987).

In what follows we shall discuss resampling methods and apply them to the computation of $\xi(r)$ for the Abell and CfA catalogues. We also discuss the presence of negative correlations between

*The bootstrap scheme cannot give any quantitative measure of systematic contributions to the errors, nor will it mirror uncertainties introduced because the data set being used is unrepresentative in some way; for example, in the problem of rich cluster clustering, by being so underpopulated with rich clusters that it cannot be regarded as a fair sample of the Universe from which to draw statistical conclusions about the whole.

Abell clusters; models with subrandom power spectra on large scales, such as the cold dark matter model discussed above, predict that ξ_c should go negative on some scale.

2 Resampling methods

The purpose of resampling a data set is to generate further data sets with a population distribution that is identical to the original true data set. This can be done by an appropriately small perturbation of the true data set. Clearly, if the perturbation is too strong we risk producing data sets with different distributions. Thus, the resampling procedure must be chosen carefully.

The optimal procedures to follow in generating pseudo data sets has been discussed in detail by statisticians (Rey 1980; Rocke & Downs 1981). We shall use the so-called 'bootstrap' resampling method whereby pseudo data sets are generated by sampling N points with replacement from the true data set of N points. It should be noted that these techniques are only designed to estimate the internal variance of the true data set. Mean values calculated over an ensemble of pseudo data sets are not expected to be good estimators of the true mean.

More precisely, suppose the underlying (unknown) probability distribution of galaxies or clusters is denoted by $P(X_1, X_2, \ldots X_n)$. We have observed the particular realization $X_1=x_1$, $X_2=x_2, \ldots, X_n=x_n$ which constitutes the catalogue under study and P^+ the observed distribution obtained when equal weight $1/n$ is placed on each of the x_i. A *bootstrap sample* is formed by taking a random sample of size n with replacement from the observed catalogue $\{x_1, x_2, \ldots x_n\}$. This operation is repeated N times to produce N bootstrap samples y_i^*, $i=1, 2, \ldots N$. We can now calculate the statistic of interest [e.g. in this case the two-point correlation function, $\xi(r)$] for each of the bootstrap samples. We label them $\xi_i^* \equiv \xi_i(y_i^*)$ where $i=1, 2, \ldots N$. The bootstrap estimate of the standard deviation of ξ_i^* is

$$\sigma_N=\left\{\sum_{i=1}^{N}[\xi_i^*-\xi^*(-)]^2/(N-1)\right\}^{1/2}, \tag{2}$$

where

$$\xi^*(-)=\sum_{i=1}^{N}\xi_i^*/N. \tag{3}$$

As $N\to\infty$, σ_N approaches a limit σ^+, which we call the bootstrap estimate of the standard error, that is

$$\operatorname*{Lt}_{N\to\infty}\ \sigma_N=\sigma^+\equiv\sigma(P^+), \tag{4}$$

providing the bootstrap samples have the same size as the original sample. The quantity σ^+ is the non-parametric maximum-likelihood estimate of the true standard error $\sigma(P)$ of the underlying distribution (Kiefer & Wolfowitz 1956) from which the galaxy catalogue was drawn by observation.

In view of the limit in (4) it is clearly best to take as many bootstrap resamples as possible. We used $N=20$ for computations of the angular correlation function and $N=100$ for the spatial correlations and $J_3(R)$ and found rapid convergence. Some idea of the rate of convergence of σ_N to σ^+ as $N\to\infty$ can be obtained from the formula derived by Efron & Tibshirani (1986),

$$X(\sigma_N)\approx\{X(\sigma^+)^2+[E(\delta_N)+2]/4N\}^{1/2}, \tag{5}$$

where $X(\sigma)=\sigma/\mu_1$ is Pearson's coefficient of variation, $E(\delta_N)$ is the expectation value of the kurtosis, $\delta_N=\mu_4/\sigma_N^4$, of the bootstrap resamples $\xi^*(-)$ and μ_i are the ith moments of the bootstrap sample distributions (see, for example, Kendall & Stuart 1976).

3 The samples

We examine samples derived from Abell's (1958) catalogue of rich clusters of galaxies and the CfA redshift survey of Davis *et al.* (1983). The Abell sample contains 104 clusters and is described in Bahcall & Soneira (1983). This data set includes all clusters located at high galactic latitude (as specified in Abell's table 1 plus the requirement $b^{II}>30°$) that have distance class $D\leqslant 4$ and richness class $R\geqslant 1$. Redshifts are available for all but one of these clusters (Hoessel, Gunn & Thuan 1980). We use the redshift estimate of 0.083 made by Bahcall & Soneira for the missing redshift of cluster A415. We shall also use a much larger subset of the Abell catalogue to calculate the angular two-point correlation function. This sample of 1664 clusters fulfils the geometrical selection criterion given above. It includes richness classes $R\geqslant 1$ and distance classes $D=5+6$. The subset of the CfA catalogue which we use is a volume-limited sample containing 489 galaxies in the north galactic cap ($b^{II}>40°$) with a maximum radial distance of $40\,h^{-1}$ Mpc. All positions are corrected for Virgocentric infall.

4 The angular correlation function

We derive the angular two-point correlation function for the Abell $D=5+6$ sample and then repeat the computation for pseudo data sets derived from it by the bootstrap procedure. The correlation function $\omega(\theta)$ was determined from the standard estimator

$$\omega(\theta)=[N_0(\theta)/N_R(\theta)]-1, \tag{6}$$

where $N_0(\theta)$ is the observed frequency of pairs in the sample and $N_R(\theta)$ is the frequency of pairs in a random catalogue within the same angular boundaries. We average $N_R(\theta)$ over an ensemble

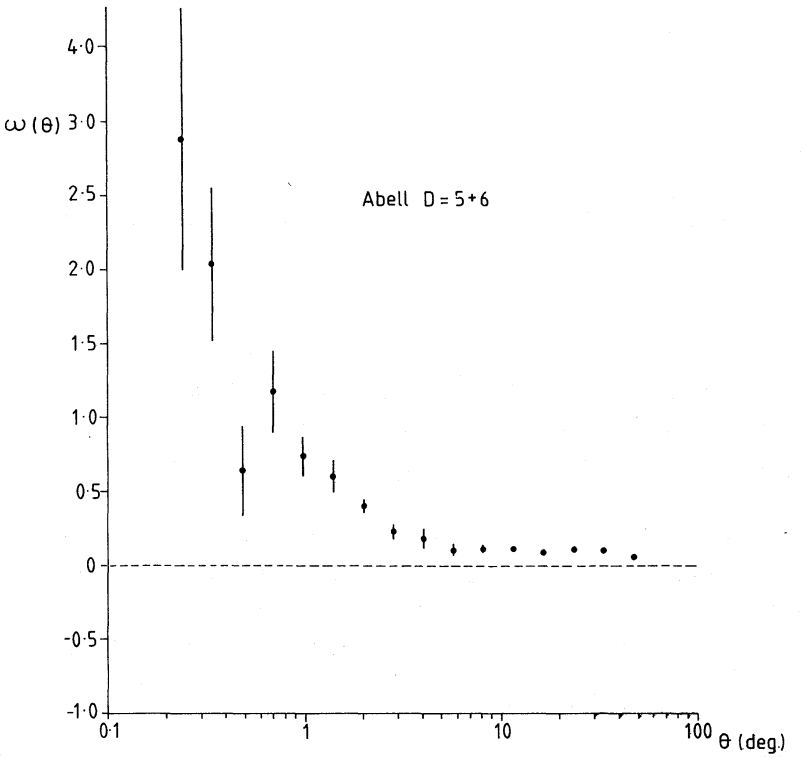

Figure 1. The two-point angular correlation function for the Abell $D=5+6$ sample. The 1σ error bars are determined from 20 pseudo data resamples of the observational data set. The correlation function remains positive at the 3σ confidence level.

of random catalogues ($\approx 10^4$ points) in order to eliminate the effects of fluctuations present in a single random realization. We ensure that each random catalogue contains the same number of points in each hemisphere as the real Abell data and the associated pseudo data sets. We fitted the correlation function with a power law

$$\omega(\theta)=(\theta/\theta_0)^{\gamma}; \qquad \theta_0, \gamma \text{ constants.} \tag{7}$$

The results are displayed in Fig. 1 and agree well with those of Bahcall & Soneira (1983). The bootstrap technique was then applied to generate 20 pseudo data sets from the original Abell data set. The standard errors in $\omega(\theta)$ for each bin over the ensemble are indicated by the error bars shown on Fig. 1. These errors are the percentile range at the 68 per cent level either side of the median (analogous to the standard deviation in Gaussian statistics). Small positive correlations, $\omega(\theta)\approx 0.1$, are present at $\theta > 40°$. The error estimates show $\omega(\theta)$ to be at least 3σ above zero.

5 The spatial correlation function

We now estimate the two-point spatial correlation functions for the Abell redshift sample and the CfA data. The above-mentioned angular criteria were adhered to when generating random data. In addition we took account in the random catalogues of the density dependence on redshift in the Abell sample by randomly selecting redshifts from the same selection function as the real data.

The resulting correlation functions were fitted by (1) and the results are displayed in Fig. 2. The values of the galaxy–galaxy and cluster–cluster correlation lengths, r_0, are found to be 5.2 and $21.9\,h^{-1}$ Mpc, with corresponding γ values of -1.87 and -1.52, respectively. Again, these values agree well with the results of previous computations.

The errors on $\xi(r)$ were estimated from 100 pseudo data sets constructed from the Abell and CfA samples. The 68 and 95 percentile error ranges in r_0 were found to be $(+0.22, -0.55)$ Mpc and $(+0.51, -0.64)$ Mpc for the CfA data and $(+7.18, -5.14)$ Mpc and $(+17.3, -9.12)$ Mpc for the Abell data. These are shown as horizontal error bars in Fig. 2.

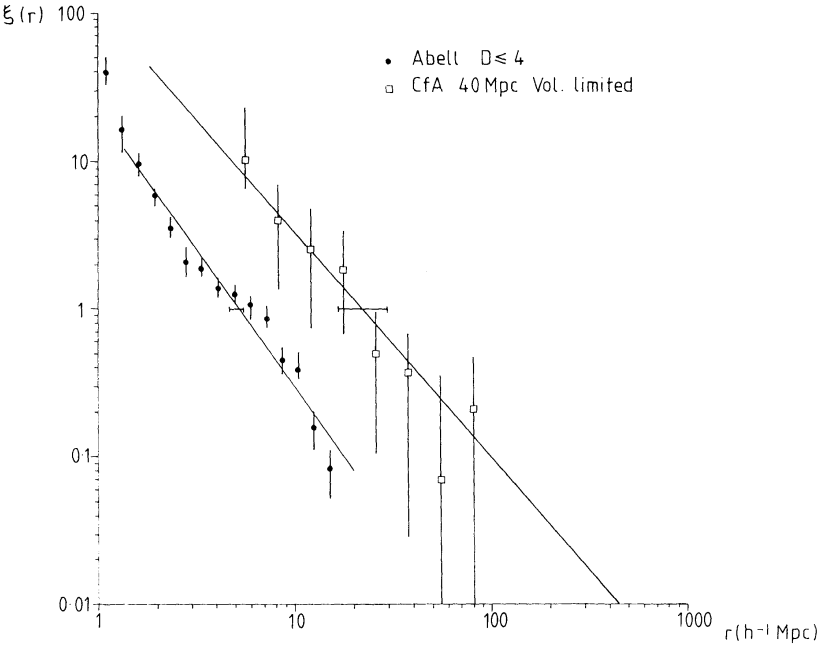

Figure 2. The spatial correlation function for the CfA and Abell data sets described in the text. The 1σ error bars are determined from 100 pseudo data resamples of the observational data set. The 1σ errors in the correlation lengths are indicated as horizontal bars.

Table 1. Computations of $J_3(R)$ for the CfA and Abell samples described in the text. The 68 per cent errors are those arising from the spread over 100 pseudo data sets generated by the bootstrap resampling procedure.

R/h^{-1} Mpc	CfA galaxies $J_3(R)/h^{-3}$ Mpc^3	68 per cent	Abell clusters $J_3(R)/h^{-3}$ Mpc^3	68 per cent
3	63	+12, −8		
5	125	+13, −19	1147	+1015, −571
10	278	+47, −28	2044	+1760, −630
15	419	+37, −71	3832	+4355, −1661
20			7898	+4930, −4978
30			11350	+6085, −6059
45			19490	+10780, −14720

We have also calculated the integral J_3 defined as,

$$J_3(R)=\int_0^R \xi(r)r^2\,dr. \tag{8}$$

Values of $J_3(R)$ and 68 percentile error ranges for both the CfA and Abell cluster samples are given in Table 1. For the latter sample J_3 is very uncertain beyond 10 Mpc.

6 Discussion

By employing the bootstrap resampling technique we have derived an estimate of the standard error associated with determinations of the two-point correlation function of cluster–cluster clustering. We find that the standard errors in $\xi(r)$ are considerably larger than the 'Poisson errors' given by Bahcall & Soneira (1983). Smaller errors are found in the calculations of the galaxy–galaxy correlations because the CfA catalogue is larger and complete. Our estimates for the error in r_0 do indicate a significant enhancement of the cluster–cluster correlation function $\xi_c(r)$ over $\xi_g(r)$ at the 3.2σ level, not withstanding the larger errors in $\xi_c(r)$. Our best estimate for the ratio of the correlation length of Abell clusters of richness class $R\geqslant 1$ and distance class $D\leqslant 4$ to the correlation length of CfA galaxies is 4.2 with a 68 percentile error range of (+1.4, −1.0). The angular correlation function of Abell clusters remains positive at least out to an angular separation $\theta=40°$. This behaviour appears also in the spatial correlations of the $D\leqslant 4$ sample but with less statistical significance.

We conclude that the enhanced amplitude of cluster–cluster clustering is statistically significant in the presence of bootstrap resampling errors. These do not include possible systematic biases in Abell's cluster catalogue nor do they include errors introduced if the samples used are unrepresentative of the Universe. The present data are inconsistent at the '2σ' level with the predictions of the standard cold dark matter models and indeed with any other model that predicts anti-clustering on large scales. Larger and more homogeneous cluster catalogues are required to confirm the observational result.

Acknowledgments

ENL was supported by a SERC post-graduate studentship and CSF was supported by a SERC post-doctoral fellowship at the Astronomy Centre, Sussex, whilst this work was performed.

References

Abell, G. O., 1958. *Astrophys. J. Suppl.*, **3,** 11.
Bahcall, N. A. & Soneira, R. M., 1983. *Astrophys. J.*, **270,** 20.

Bardeen, J., Bond, J. R., Kaiser, N. & Szalay, A., 1986. *Astrophys. J.,* **304,** 15.
Barnes, J., Dekel, A., Efstathiou, G. & Frenk, C. S., 1985. *Astrophys. J.,* **295,** 368.
Barrow, J. D. & Bhavsar, S. P., 1987. *Q. Jl R. astr. Soc.,* **28,** 000.
Barrow, J. D., Bhavsar, S. P. & Sonoda, D. H., 1984. *Mon. Not. R. astr. Soc.,* **210,** 19P.
Blumenthal, G. R., Faber, S. M., Primack, J. R. & Rees, M. J., 1984. *Nature,* **311,** 517.
Coles, P., 1986. *Mon. Not. R. astr. Soc.,* **222,** 9P.
Davis, M., Efstathiou, G., Frenk, C. S. & White, S. D. M., 1985. *Astrophys. J.,* **292,** 371.
Davis, M., Huchra, J., Latham, D. & Tonry, J., 1983. *Astrophys. J.Suppl.,* **52,** 89.
Davis, M. & Peebles, P. J. E., 1983. *Astrophys. J.,* **267,** 465.
Efron, B. & Tibshirani, R., 1986. *Stat. Sci.,* **1,** 54.
Hoessel, J. G., Gunn, J. E. & Thuan, T. X., 1980. *Astrophys. J.,* **241,** 486.
Kaiser, N., 1984. *Astrophys. J.,* **284,** L9.
Kaiser, N., 1986. *Mon. Not. R. astr. Soc.,* **219,** 785.
Kendall, M. & Stuart, A., 1976. *The Advanced Theory of Statistics,* Vol. 1, 4th edn, Charles Griffen & Co., London.
Kiefer, J. & Wolfowitz, J., 1956. *Ann. Math. Stat.,* **27,** 887.
Klypin, A. A. & Kopylov, A. I., 1983. *Soviet Astr. Lett.,* **9,** 75.
Lucey, J. R., 1981. *DPhil thesis,* University of Sussex.
Lucey, J. R., 1983. *Mon. Not. R. astr. Soc.,* **204,** 33.
Peacock, J. A. & Heavens, A. F., 1985. *Mon. Not. R. astr. Soc.,* **217,** 805.
Peebles, P. J. E., 1973. *Astrophys. J.,* **185,** 423.
Postman, M., Geller, M. J. & Huchra, J. P., 1986. *Astr. J.,* **91,** 1267.
Rey, W., 1983. *Introduction to Robust and Quasi-Robust Statistical Methods,* Springer–Verlag, Berlin.
Rocke, D. M. & Downs, G. W., 1981. *Commun. Stat. Sim. Comp.,* **B-10,** 221.
Turok, N., 1985. *Phys. Rev. Lett.,* **55,** 1801.
White, S. D. M., Frenk, C. S., Davis, M. & Efstathiou, G., 1986. Preprint.