Mon. Not. R. astr. Soc. (1982) **201**, 867–883

The estimation of galaxy angular correlation functions

Paul C. Hewett *Department of Astronomy, University of Edinburgh, Blackford Hill, Edinburgh, EH9 3HJ, Scotland*

Received 1982 March 24; in original form 1981 September 9

Summary. Analysis of galaxy samples shows that significant bias in the estimation of galaxy correlation functions can occur depending on the type of estimator and edge correction applied to the sample. A general technique for calculating estimates of correlation functions that is free from systematic errors is described. The technique is applicable to correlation estimates for many types of astronomical data and is strongly recommended for future use.

A sharp break in the power-law form of the covariance function at scales of $7H^{-1}$ Mpc is found in two galaxy samples confirming the results from the Lick counts obtained by Groth & Peebles. Some evidence for weak anti-clustering at large scales is found.

1 Introduction

The two-point and, more recently, higher order correlation functions (Peebles 1973; Peebles & Groth 1975; Fry & Peebles 1978) have become the predominant form of clustering analysis applied to galaxies. Although the Princeton group and others (Peebles & Groth 1975; Fry & Peebles 1978; Bonometto & Lucchin 1980) have pioneered the estimation of higher order functions, much recent work deals with the covariance function (e.g. Dautcourt *et al.* 1978; Shanks *et al.* 1980a, henceforth SFEM). Correlation techniques are also now finding uses in all fields of clustering analyses; clustering of radio sources (Masson 1979; Seldner & Peebles 1981), environments of radio galaxies (Longair & Seldner 1979), three-dimensional clustering of quasars (Osmer 1982), distribution of quasar absorption lines (Sargent *et al.* 1980), distribution of galactic stars (Shanks, Phillipps & Fong 1980b; Bahcall & Soneira 1981) and many others. Correlation functions are now one of the major clustering techniques applied to astronomical data.

Considering galaxy clustering studies in particular, there are considerable discrepancies between the Princeton group's results and those of more recent studies. Results from machine measures of deep galaxy samples recently obtained suggest that the covariance function shows large variations from field to field. Evidence for anticlustering, deviations from the accepted power-law relation for the covariance function at very small angular scales and significant correlations extending to beyond $10H^{-1}$ Mpc are evident in the machine

measures. It is shown in this paper that these discrepancies are partially due to the different analysis techniques employed. The discrepancies are discussed in Section 2. In Section 3 the estimation of $W(\theta)$ is discussed in some detail and a correction procedure first suggested by Sharp (1979) is further developed to reduce the estimators to a self-consistent form. A new galaxy sample is presented in Section 4. In Section 5, three galaxy samples are analysed and the discrepancies between estimator/edge correction combinations illustrated, before the correction procedure is applied. Section 6 contains a discussion of some of the features present in the corrected estimates of $W(\theta)$.

2 The observations

The Zwicky, Lick and Jagellonian analyses (Peebles & Hauser 1974; Groth & Peebles 1977, henceforth GP 77; Peebles 1975) demonstrated the consistency of the covariance function estimates over a large range of sample depth. However, considerable difficulties were encountered at large angular scales in separating effects due to galactic obscuration, variable catalogue depth and changing selection effects from intrinsic clustering. These effects combined with relatively poor plate material – by modern standards – effectively limit the information obtainable from these early catalogues. In particular no reliable estimates of the sign or amplitude of the spatial correlation function $\xi(r)$ exist beyond $15H^{-1}$ Mpc (Peebles 1980, section 57). Throughout, H denotes the value of the Hubble constant in units of $100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

A power-law model of the form

$$W(\theta) = A\,\theta^{-\delta} \tag{1}$$

is found to fit the observations at small angular scales with the power-law index $\delta \sim 0.8$, (GP 77). The amplitude, A, which is a strong function of sample depth (due to the large increase in the number of overlapping structures appearing in projection) generally scales well, though there are some discrepancies (GP 77; section V). In the analysis of the most recent reduction of the Lick counts (Seldner *et al.* 1977; GP 77) a steep drop from this power law at a scale $9H^{-1}$ Mpc is found, with a somewhat weaker feature indicated in the Zwicky catalogue (GP 77). No significantly negative values of $\hat{W}(\theta)$ were found in any of the three samples.

Large new catalogues of faint galaxies have been difficult to obtain and the only deep wide-angle samples have come from the United Kingdom Schmidt Telescope (UKSTU)/COSMOS combination at the Royal Observatory Edinburgh. These samples were subsequently analysed by groups at Durham and Edinburgh (MacGillivray & Dodd 1979, henceforth MD 79; Dodd & MacGillivray 1980, henceforth DM 80; SFEM). The samples of MD 79, DM 80 and SFEM total eight plates of six fields, each with an area of ~ 15 square degrees, corresponding to a linear size of $\sim 40H^{-1}$ Mpc at the typical sample depth of $Z^* \sim 0.2$–0.25. With suitable scaling and allowance for the somewhat different selection functions involved these should be directly comparable to earlier work.

The covariance function estimates from these deeper samples generally follow the familiar power-law relation at small angular scales with the constants A and δ in reasonable agreement with the shallower samples, although the agreement is not totally compelling (see section 5 of SFEM for a discussion). At large angular scales the shape of $\hat{W}(\theta)$ shows substantial variations: the break point from the power law can be at much smaller projected scales (e.g. $3H^{-1}$ Mpc, SFEM) or be non-existent (e.g. fig. 2a of MD 79) and significant negative values of $\hat{W}(\theta)$ are sometimes present (e.g. fig. 2b of MD 79 and fig. 2 of DM 80). Substantial discrepancies are present between shallow and deep samples at large angular scales. These are particularly serious theoretically as the existence of the power-law break point may indicate

the scale at which the change from non-linear to linear growth of clustering occurs (Davis, Groth & Peebles 1977) and the failure of the covariance function to become negative at larger scales ($>15H^{-1}$ Mpc) would show that clusters are not surrounded by regions of decreased galaxy density – or 'holes' – (Peebles 1974). The discrepancies require explanation because of the apparently high quality of the machine data.

Four possible explanations for the discrepancies suggest themselves.

(a) The intrinsic clustering properties of the galaxies in the shallow and deep samples are significantly different.

(b) Insufficient allowance has been made for the differences between the selection functions applying for shallow and deep samples, leading to a serious error in the scaling relations.

(c) The different analysis techniques used to derive the estimators from the shallow and deep samples are producing the disagreement.

(d) There are still systematic errors present in the COSMOS samples.

Considering each in turn: (a) is extremely unlikely as the disagreement in the power-law break scale persists when the Jagellonian sample is compared to the brightest SFEM sample which is of shallower depth. Furthermore, the lookback time difference between $m_B \sim 20.3$ (Jagellonian) and $m_B \sim 21.5$ (SFEM) is far too short for significant clustering evolution to occur. (b) is also ruled out by the Jagellonian/SFEM comparison; judicious changes in scaling to obtain agreement between the breakpoints would result in the derived amplitudes A being in conflict. Further, detailed modelling of the selection functions for deep samples has been carried out by the Durham group (Ellis, Fong & Phillipps 1977; Phillipps *et al.* 1978; SFEM) and their results have been independently confirmed as part of this investigation. (c) the analysis techniques employed on the deep samples are different from those used on the shallow samples and I show below how the different analysis techniques give rise to some of the observed discrepancies. The Durham group have taken considerable trouble over their estimation procedures and although the use of the filtering technique as in SFEM is not found to be desirable (see Section 5 below) the estimation procedure differences are not the prime cause of the discrepancies in SFEMs case. (d) recent work at Edinburgh indicates that some aspects of the data obtained from deep J and F plates are unsatisfactory. In particular the high success rates claimed for star–galaxy separation appear to be optimistic in some cases. Systematic variations in image parameters over plates have been found in COSMOS data, with these variations being correlated with the large sky background variations often present on particular UKSTU J and F plates. For this reason new data from four V plates of one of the South Galactic Pole (SGP) fields analysed by SFEM are used. The results described in Section 5 show considerable disagreement with those of SFEM and support the view that at least part of the remaining discrepancies are due to systematic effects in some of the earlier COSMOS samples.

3 Estimation of $W(\theta)$

Sharp (1979) has considered some of the problems of edge correction when estimating $W(\theta)$ from a bounded sample. I shall emphasize the effects of differing estimator definitions as well as the edge correction effects.

3.1 DEFINITIONS OF THE ESTIMATOR $\hat{W}(\theta)$

The covariance function $W(\theta)$ is defined by

$$\delta P = N^2[1 + W(\theta)]\delta\Omega_1\delta\Omega_2 \qquad (2)$$

where δP is the joint probability of finding one object in solid angle element $\delta\Omega_1$ and another in solid angle element $\delta\Omega_2$ separated by an angle θ. N is the mean surface density of the sample.

When a sample contains angular positions of objects an estimate of $W(\theta)$ is

$$\hat{W}(\theta) = \frac{N_p(\theta)}{N_c N \delta\Omega} - 1. \tag{3}$$

$N_p(\theta)$ is the number of distinct pairs between $\theta - \Delta\theta/2$ and $\theta + \Delta\theta/2$. N_c is the number of objects used as centres. N is the mean surface density of the sample and $\delta\Omega$ is the solid angle of the ring radius θ thickness $\Delta\theta$.

The equivalent Monte Carlo estimator is

$$\hat{W}(\theta) = \frac{N_p(\theta)}{N_r(\theta)} - 1 \tag{4}$$

where $N_p(\theta)$ is the number of pairs in the sample with separations between $\theta - \Delta\theta/2$ and $\theta + \Delta\theta/2$, $N_r(\theta)$ the number of pairs for the same number of objects distributed randomly over an identical area. This method removes the need for separate calculation of edge corrections.

For a sample where counts N_i are obtained in cells, one estimator (e.g. SFEM) is

$$\hat{W}(\theta) = \frac{\langle N_i N_j \rangle}{N^2} - 1 \tag{5}$$

where the angular brackets $\langle\ \rangle$ denote averages over all cells with separations between $\theta - \Delta\theta/2$ and $\theta + \Delta\theta/2$. N is the mean surface density of the sample.

An important point is that the estimate of N in the definition of $W(\theta)$ – equation (2) – is derived from the total number of objects in the sample since the mean sample density is used to normalize the estimators. If, as in practice, $W(\theta)$ is positive at small angular scales then it must become negative at some larger scale to satisfy the integral constraint operating on the estimator (see e.g. Peebles 1980, section 32). However, the low amplitude of $W(\theta)$ in typical two-dimensional galaxy samples means the effect is small in practice. For want of better terminology I shall term the estimators described above as 'direct'.

SFEM proposed a modification to the direct estimator by employing a filtering technique (SFEM, section 3.3). This technique was applied to allow for large-scale gradients in the galaxy distribution thought to be due to patchy galactic obscuration. A moving average filter is applied to the data, the estimator is calculated according to equation (5) and the count in the i–jth cell N_{ij} is replaced by the value

$$N'_{ij} = N N_{ij}/N_s \tag{6}$$

where N is the mean count for the entire sample and N_s the mean count for the immediately surrounding area. The size of this surrounding area used to calculate N_s determines the filter scale length. SFEM used a filter length equal to half the sample size.

An 'ensemble' estimator

$$\hat{W}(\theta) = \frac{\langle N_i\, N_j \rangle}{\langle N_i \rangle \langle N_j \rangle} - 1 \tag{7}$$

has been employed by the Princeton group (Peebles 1975), together with slight variations (Peebles & Hauser 1974)

$$\hat{W}(\theta) = \frac{\langle N_i\, N_j \rangle}{\langle (N_i + N_j)/2 \rangle^2} - 1 \tag{8}$$

and (Dautcourt *et al.* 1978)

$$\hat{W}(\theta) = \frac{\langle (N_i - N)(N_j - N) \rangle}{N^2}. \tag{9}$$

Note the completely different approach to normalization in the ensemble estimators: the normalization is derived only from the regions of the sample used to calculate $\hat{W}(\theta)$ at each scale and the integral constraint does not apply to this type of estimator. For inhomogeneous or anisotropic samples – reflecting either large-scale structure in the galaxy distribution or some external factors – it is possible that $\langle N_i \rangle \neq \langle N_j \rangle \neq N$ at one or more scales θ, consequently the resulting estimates of $W(\theta)$ will be different. The exact relation between the differing estimator of $W(\theta)$ may be seen by putting $\langle N_i \rangle = N + \epsilon_1$ and $\langle N_j \rangle = N + \epsilon_2$. As a simple illustration imagine a one-dimensional array of 10 cells with a strong density gradient increasing from left to right superimposed on any clustering present. When the covariance function is calculated at scale 'one' – i.e. between adjacent cells – then the leftmost nine cells will contribute to $\langle N_i \rangle$ and the rightmost nine to $\langle N_j \rangle$ – vice versa if we start at the right-hand edge when calculating $\hat{W}(\theta)$. Due to the density gradient $\langle N_i \rangle < N$ and $\langle N_j \rangle > N$, so $\langle N_i \rangle \neq \langle N_j \rangle \neq N$, where N is the mean count per cell.

3.2 EDGE CORRECTION

Allowance must be made for the sample boundaries when estimating $W(\theta)$. Three techniques have been extensively used:

(A) The number of pairs is scaled by considering the fraction of each ring (or equivalently the number of cell pairs) within the sample boundaries (this technique is favoured by the Princeton group).

(B) Only objects with distance at least $\theta_{\max}$ from the sample edges are used as centres, where $\theta_{\max}$ is the largest scale examined.

(C) Only objects with distance at least θ from the sample edges are used as centres, where θ is the scale being examined.

The major advantage of technique (A) is the use of the maximum amount of data in the calculation of $\hat{W}(\theta)$ at each scale, thereby ensuring that statistical noise is reduced to a minimum. The use of as much of the data as possible to calculate $\hat{W}(\theta)$ also means that the method is least susceptible to systematic effects due to inhomogeneities and anisotropies in the data.

Methods (B) and (C) have been used in combination by Phillipps *et al.* (1978), MD 79 and DM 80. The fraction of the data used to calculate $\hat{W}(\theta)$ at each scale is much smaller than in method (A) so the noise is increased. Further, the methods are sensitive to inhomogeneities in the data as the fraction of the data used to calculate $\hat{W}(\theta)$ changes with scale. In principle, there should be no difference in which method of edge correction is used provided the samples are homogeneous and isotropic, which would ensure that any reasonable subsample of the data will be a fair representation of the whole sample.

When considering samples as cell counts below, correction methods (B) and (C) are applied by considering only cell pairs where one or both are at least at a distance of $\theta_{\max}$ or

θ respectively from the boundaries. The calculations on the galaxy samples in Section 5 are only carried out to a scale one quarter of the sample size for methods (B) and (C) because of the small fraction of the data contributing to $\hat{W}(\theta)$. Throughout the paper the calculation of the noise associated with $\hat{W}(\theta)$ is made according to

$$[\delta \hat{W}(\theta)]^2 = [1 + \hat{W}(\theta)]/N_p \tag{10}$$

where N_p is the number of distinct pairs contributing to the calculation of $\hat{W}(\theta)$ at each scale. Sharp (1979) has shown that this is a slight overestimate in terms of 'random statistical error'. The 2σ error bars thus calculated, shown in the figures, are representative of those at neighbouring points which have been omitted for reasons of space. Data points have been joined up in the figures for clarity but no 'interpolation' between data points is implied.

3.3 CROSS CORRELATION OF GALAXY AND RANDOM SAMPLES

Sharp (1979) suggested a method for investigating bias in the estimators of $W(\theta)$ when data is available as individual positions. The technique may be understood as follows: any homogeneous and isotropic sample of data should be unrelated to a set of points randomly distributed over the identical sample area. Consequently the cross-correlation function between the data and random samples should be zero at all scales.

The cross-correlation function is defined by

$$\delta P = N_1 N_2 [1 + W_{12}(\theta)] \delta\Omega_1 \delta\Omega_2 \tag{11}$$

where δP is the joint probability of finding an object of type '1' in solid angle element $\delta\Omega_1$ and an object of type '2' in solid angle element $\delta\Omega_2$. The surface densities N_1 and N_2 of both samples enter into the equation and $W_{12}(\theta)$ is a measure of the cross-correlation of the samples, $W_{12}(\theta) = 0$ for unrelated samples. Note the definition is symmetric $W_{12}(\theta) = W_{21}(\theta)$. Considering data as individual positions, then in practice $\hat{W}_{\text{galaxy/random}}$, $(\hat{W}_{\text{g/r}})$ will always be zero no matter how inhomogeneous the galaxy distribution, as the random sample is always randomly distributed about each galaxy – $\hat{W}_{\text{g/r}}$ is calculated by performing a sum over all the galaxies. In contrast choosing a random point as centre and summing over all the random data to calculate $\hat{W}_{\text{random/galaxy}}$ $(\hat{W}_{\text{r/g}})$ may give a non-zero result for an inhomogeneous or anisotropic galaxy distribution. This is equivalent to saying that the value of $\hat{W}_{\text{r/g}}$ is at least partially determined by the positions of the galaxies relative to the sample boundaries, e.g. when a large cluster is at the centre of the field, (*cf.* fig. 5 of Sharp 1979).

For counts tabulated in cells then $\hat{W}_{\text{g/r}}$ and $\hat{W}_{\text{r/g}}$ are no longer distinct. The estimators $\hat{W}_{\text{r/g}}$ and $\hat{W}_{\text{g/r}}$ are calculated by summing over both sets of objects so that $\hat{W}_{\text{g/r}}$ is not necessarily equal to zero. The covariance function $\hat{W}(\theta)$ can therefore be corrected by subtracting the contributions of $\hat{W}_{\text{r/g}}(\theta)$ and $\hat{W}_{\text{g/r}}(\theta)$ from the original estimator. A general equation applicable to data both as individual positions and as cell counts is

$$\hat{W}\,\text{corr}\,(\theta) = \hat{W}(\theta) - \hat{W}_{\text{g/r}}(\theta) - \hat{W}_{\text{r/g}}(\theta). \tag{12}$$

For samples as individual positions $\hat{W}_{\text{g/r}}(\theta) = 0$ and the equation is still valid. This equation is applied in Section 5 to obtain self-consistent results from different estimator/edge correction combinations.

4 The galaxy samples

Three samples were chosen for investigation.

(1) The data of MacGillivray & Dodd (1980) consisting of 28 872 objects. The area of the sample is 14.6 square degrees and full details of the sample are contained in MacGillivray &

Dodd (1980). The covariance function for the sample published in DM 80 shows large discrepancies with the Princeton results.

(2) The Jagellonian sample analysed by Peebles (1975). The sample contains 12 145 galaxies in an area of 36 square degrees. The effective limiting magnitude of the sample is $m_B \sim 20.3$ (GP 77).

(3) An SGP V sample which contains 5445 objects in an area of 4.8×4.6 degrees, with an area of 0.24 square degrees drilled out near bright images in the field where the detection of galaxies by COSMOS is impaired. The sample is centred on the SGP and covers a somewhat larger area than that of SFEMs central field. The galaxy sample was obtained in parallel with the stellar sample described by Reid & Gilmore (1981) in the same field. I shall summarize the important improvements incorporated in this sample compared to those previously obtained from COSMOS.

All images in the sample appear on at least two V plates ensuring that all spurious images are eliminated from the sample. In the past the apparent detection of 'compact galaxy clusters' around the brightest stars on J and F plates has proved a serious problem. The data is from unhypered IIa-D V band plates which exhibit very small sky background variations over the entire COSMOS measured area. These small variations, amounting to a few per cent in relative intensity, are explained by telescopic vignetting in contrast to the typically 10 per cent variations in sky background present on J and F plates. These variations on the J and F plates have been found to correlate with the large systematic shifts in COSMOS image parameters present in measures of some plates. This problem has been completely eliminated in this sample.

Star–galaxy classification on the V plates is much improved due to several factors.

(i) No shifts in COSMOS image parameters are present over the measured part of the plate so a global discriminator may be applied without the application of approximate and *ad hoc* corrections to the discrimination criterion as a function of plate position.

(ii) The lower image density on the V plates due to the brighter limiting magnitude reduces the number of multiple stellar images, which are nearly always classified as galaxies by the star/galaxy separation techniques applied to COSMOS measures. consequently the stellar contamination of the galaxy sample is much reduced.

(iii) A large number – 2500 – of images were classified from regions spanning the extremes of the measured area using glass copies of the UKSTU J survey limited at $J \sim 23$ and in one area a high quality AAT R plate limited at $R \sim 24$. Thus relatively faint images on the V plates have been classified using much deeper plate material, virtually removing the problem of obtaining accurate eye classification at faint magnitudes.

(iv) The sample has been deliberately confined to a 'bright' subsample of the full galaxy sample. The COSMOS selection limit corresponds to a stellar magnitude of $M_v \sim 18.5$. Calculation of the selection function for galaxies in the V band allowing for thresholding effects indicates this corresponds to an effective sample depth $D_{eff} \sim 340 H^{-1}$ Mpc.

(v) The star galaxy discrimination criterion has been applied so that an almost complete sample of galaxies is obtained at the expense of more contamination due to misclassified stars. There is thus no systematic rejection of the most compact galaxies – usually a difficult effect to allow for in subsequent analyses. For the SGP V sample 99 per cent galaxies are included with 8 per cent contamination from stars (both as a percentage of the total number of galaxies).

5 Results

Fig. 1 shows the covariance function estimates for the Jagellonian field using three different estimators – direct (equation 5), filtered (equation 6) and ensemble (equation 7) – all using

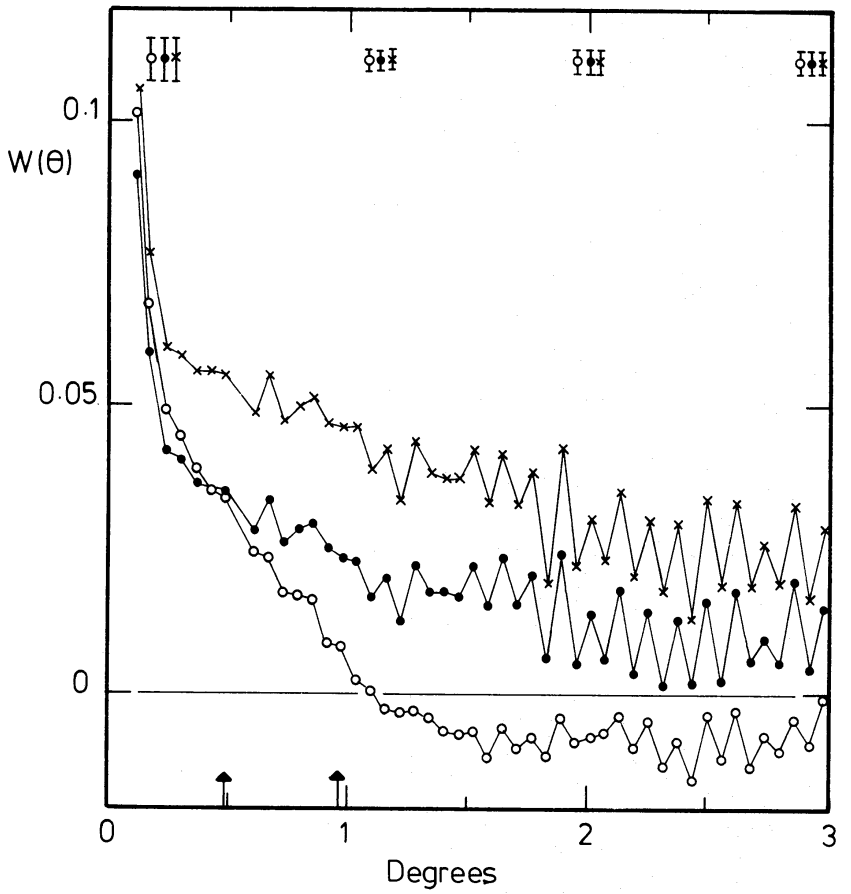

Figure 1. Covariance function estimates for the Jagellonian field using the estimators of equations (5 ×), (6 •) and (7 ○) all using edge correction method (A). Representative ± 2σ error bars are shown above the curves for each estimator. This convention for error bars is adopted throughout the paper for clarity. Note the large differences between the three types of estimator.

edge correction method (A). The filtered estimator was obtained using a filter length equal to half the sample diameter (i.e. the same scale relative to the sample size as employed by SFEM). Note the large differences evident between the ensemble, filtered and direct estimators shown in Fig. 1. For any particular sample the ensemble estimators of equations (7), (8) and (9) gave essentially identical results and equation (7) will be adopted as representative from now on. Similarly for each sample analysed the direct estimators of equations (3), (4) and (5) gave identical results within the random statistical errors and for clarity I adopt the estimator of equation (5) as representative of direct estimators.

In Fig. 2 four different estimates of $W(\theta)$ from the Jagellonian sample using the filtered estimator of equation (6) with four filter scale lengths – 50, 38, 25 and 13 per cent of the sample size – show substantial differences as expected. The derived form of $\hat{W}(\theta)$ is entirely dependent on the filter length chosen.

The effects of edge correction on the differing estimators are illustrated in Fig. 3(a), (b) and (c). Estimators have been applied to the DM 80 data using each of the three edge corrections discussed in Section 3.2. The similarity of the ensemble estimators (Fig. 3a) is obvious in comparison with the direct (Fig. 3b) and filtered (Fig. 3c) estimators which vary considerably depending on the type of edge correction used. Note that at small angular scales method (C) uses nearly all the available data and the estimates of $W(\theta)$ are similar to those of method (A). At one quarter of the sample size (the scale chosen for method B) methods

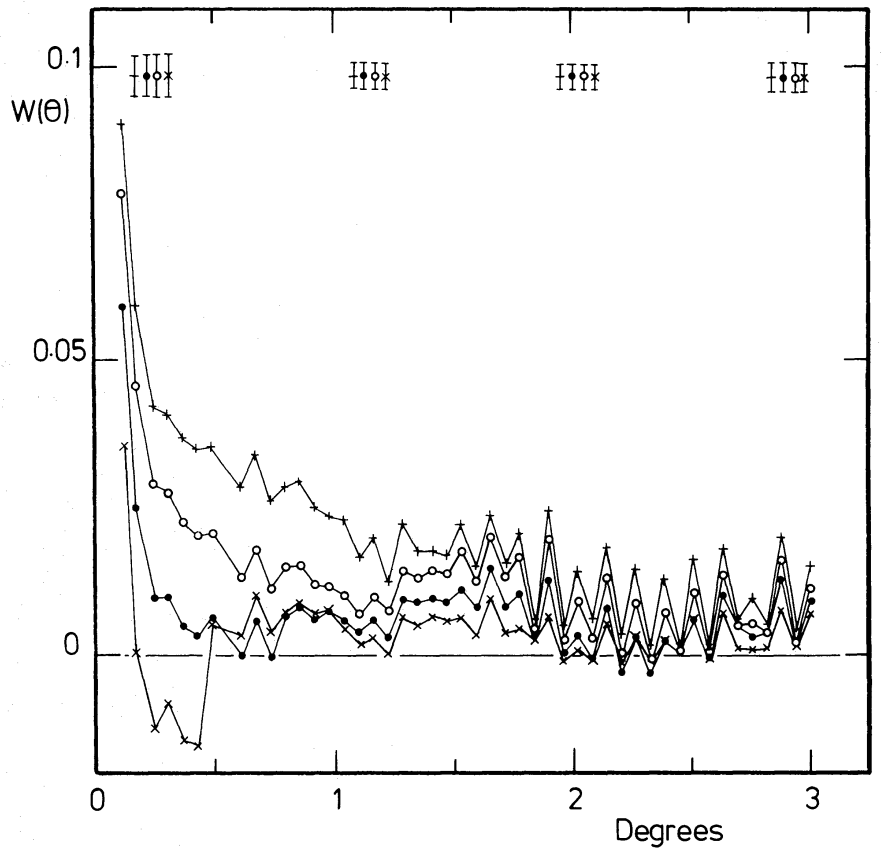

Figure 2. Four estimates of the covariance function for the Jagellonian field using the filtered estimator of equation (6) employing different filter lengths. The filter scales as percentages of the sample diameter were 50 (+), 38 (○), 25 (•) and 13 per cent (×). The results illustrate that the form of the estimators is completely dependent on the filter scale chosen. All the estimators were calculated using edge correction method (A).

(B) and (C) are using exactly the same fraction of the data and the two estimates of $W(\theta)$ are equal.

To obtain the corrected estimates of $W(\theta)$ for each of the three galaxy samples under investigation five random data sets containing the same number of objects distributed over identical areas were generated. The galaxy samples were cross-correlated with each of the appropriate random samples. The average value of the cross-correlation calculated giving estimates of $W_{r/g}$ and $W_{g/r}$ for each sample. This procedure was followed for each of the estimator/edge correction combinations. The mean values of $\hat{W}_{r/g}$ and $\hat{W}_{g/r}$ for each of the equivalent $\hat{W}(\theta)$ estimators in Fig. 3(a), (b) and (c) are shown in Figs 4(a), (b) and (c) and 5(a), (b) and (c) respectively. Values of $\hat{W}_{r/g}$ and $\hat{W}_{g/r}$ close to zero indicate that the correction to be applied to the original estimators is small. Conversely, values of $\hat{W}_{r/g}$ and $\hat{W}_{g/r}$ well away from zero indicate the corrections are large. Note that the corrections for the ensemble estimators shown in Figs 4(a) and 5(a) are equal to zero within the noise, in contrast to the large fluctuating corrections for the direct estimators (Figs 4b and 5b) and the filtered estimators (Figs 4c and 5c). The corrected estimates of $W(\theta)$ calculated according to equation (12) are shown in Fig. 6(a), (b) and (c), where the reduction to a common form is readily apparent.

The residual differences between the corrected estimates of $W(\theta)$ using the various edge corrections are due to intrinsic variations in the clustering properties of the fractions of the

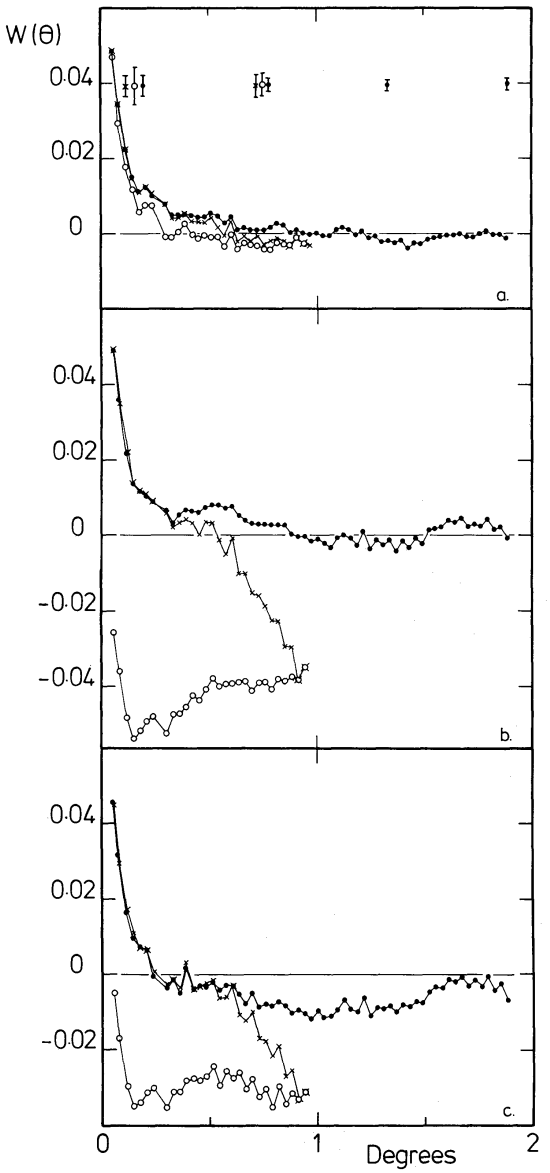

Figure 3. (a) The ensemble estimators from the DM 80 field using edge correction methods A (•), B (○) and C (×). The symbol types for the edge correction methods are maintained throughout Figs 3, 4, 5 and 6. Note the similarity between the estimators despite the different methods of edge correction. (b) Direct estimators for the DM 80 field with the three different edge corrections as in Fig. 3(a). Large differences are present showing that the estimator is strongly dependent on the type of edge correction employed. (c) Filtered estimators for the DM 80 field with the three different edge corrections as in Fig. 3(a). As with the direct estimators shown in Fig. 3(b) the estimators show large variations.

data used to calculate each estimate. The best estimate is that which uses the maximum amount of data, i.e. employs edge correction method (A). Similar reductions to a common form have been obtained for the Jagellonian and SGP V samples as well as numerous simulations. In all cases an ensemble estimator employing edge correction method (A) gives an estimate of $W(\theta)$ free of bias. No corrections are required and the effect of making the correction is merely to add noise. The same results can be obtained using direct or filtered estimators providing a correction using equation (12) is made. However, this approach necessitates more calculation and results in considerably more statistical noise. For these reasons it is recommended that the ensemble approach always be used. The procedure is easy to use

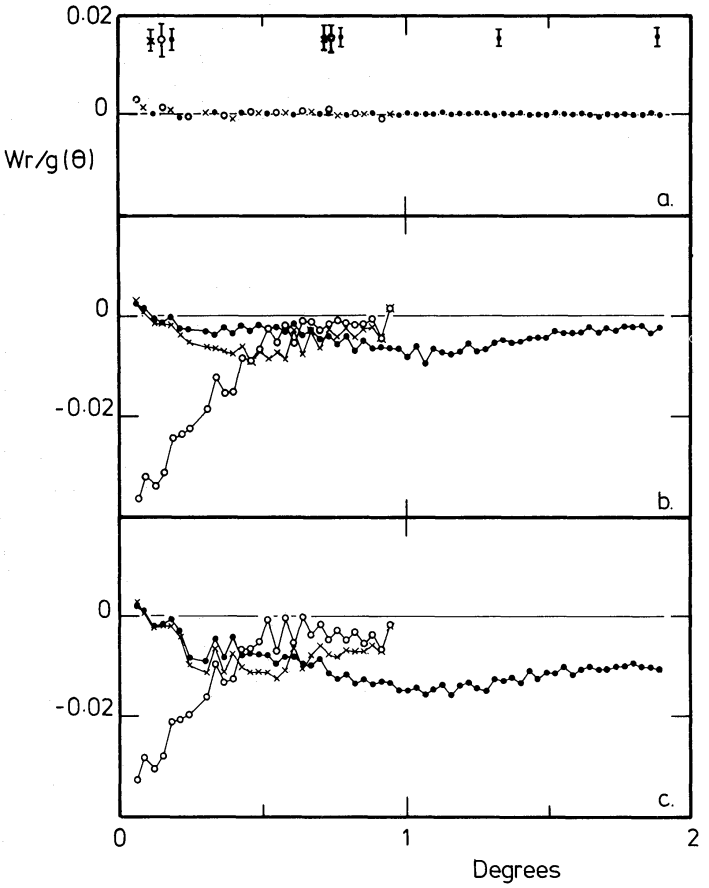

Figure 4. (a) The mean cross-correlation corrections $W_{r/g}$ for the ensemble estimators in Fig. 3(a). The cross-correlations are equal to zero within the noise. Compare the virtually non-existent corrections with those in Fig. 4(b) and (c). Symbol types are as in Fig. 3 for each type of edge correction. As the values of $W_{r/g}$ are so close to zero for the ensemble estimators only every third point for each method of edge correction has been plotted so that the individual data points are visible. (b) Cross-correlation corrections $W_{r/g}$ for the direct estimates of Fig. 3(b). Significant deviations from the zero level are evident. (c) Cross-correlation correction $W_{r/g}$ for the filtered estimates of Fig. 3(c).

and may be applied to one, two- or three-dimensional data. The procedure is particularly effective when high density contrasts of large-scale variations are evident in the data.

Fig. 7 shows the ensemble estimate of $W(\theta)$ for the SGP V sample together with the ensemble estimate of $W(\theta)$ for the parallel sample of stellar objects. This sample contains 15 145 stellar objects distributed over exactly the same area as the galaxy sample. The covariance function for the stars is completely featureless indicating the success of the star–galaxy separation and that the stars are randomly distributed on all scales. This rules out the possibility that any plate, measuring or analysis artefacts are affecting the results from the galaxy sample, unless any effect applies only to galaxies. As the sample is well away from the plate limit this is unlikely.

Fig. 8 shows the ensemble estimate of $W(\theta)$ for the DM 80 field together with the published estimate, which has been replotted on a linear scale.

The Jagellonian data shows a strong feature at angular scale ~ 1.0 degrees corresponding to a linear scale of $6–7H^{-1}$Mpc at the derived sample depth of $383H^{-1}$Mpc (GP 77). Beyond this scale the covariance function maintains small negative values until the limit set by the sample size, which corresponds to $18H^{-1}$Mpc. No complex corrections were performed on the Jagellonian data to remove large-scale inhomogeneities as was necessary

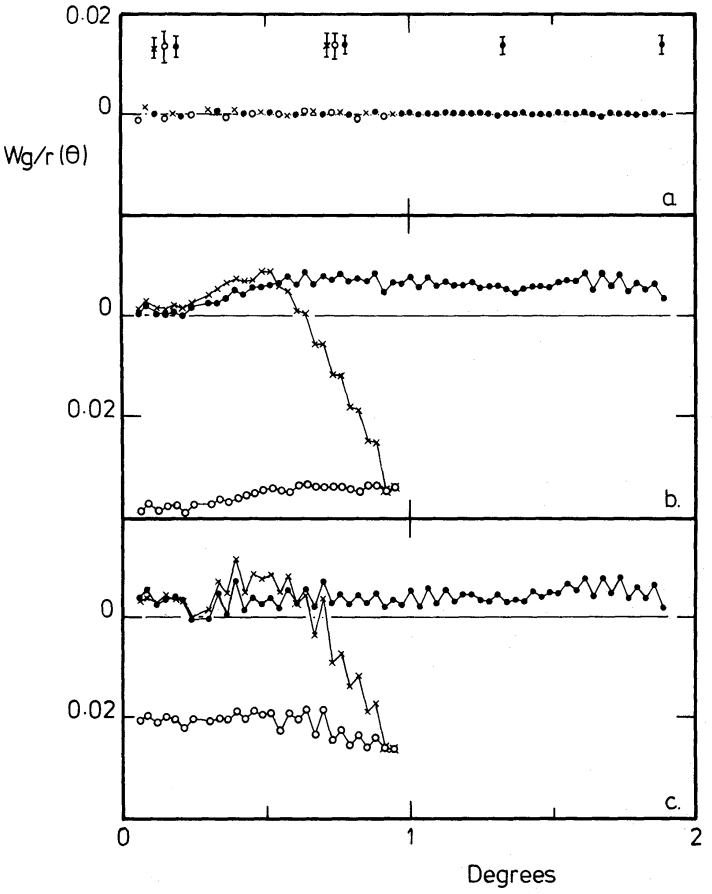

Figure 5. (a) The mean cross-correlation corrections $W_{g/r}$ for the ensemble estimators in Fig. 3(a). The cross-correlation corrections are equal to zero within the noise. As in Fig. 4(a) the cross-correlations for the ensemble estimators are so close to zero that only every third point has been plotted so that individual data points are visible. (b) Cross-correlations $W_{g/r}$ for the direct estimates of Fig. 3(b). (c) Cross-correlation $W_{g/r}$ for the filtered estimates of Fig. 3(c).

for the Lick sample. This supports the view that the break is not an artifact of the correction procedures used by Groth & Peebles.

The covariance function for the SGP V sample is positive to ~ 1.4 degrees and then drops away to low-amplitude negative values, similar to the Jagellonian sample. The coefficients A and δ fitted to the data up to a scale of 1.3 degrees are $\delta = 0.89 \pm 0.12$, $A = 0.13 \pm 0.04$. Allowing for the somewhat different COSMOS selection function compared to the earlier samples and the absence of any variable galactic extinction in the field the values of A and δ agree well with the results from the Lick and Jagellonian samples. The position of the power-law break point in the SGP sample is $8H^{-1}$ Mpc, also in good agreement with the Lick and Jagellonian samples.

6 Discussion

The previous sections have shown that large discrepancies are present in the published estimates of $W(\theta)$ and that these may in part be ascribed to the type of estimator edge correction employed. The difference between direct and filtered estimators and the ensemble estimators originally used by the Princeton group has not been generally appreciated. This difference is important – the integral constraint implicit in the direct method results in

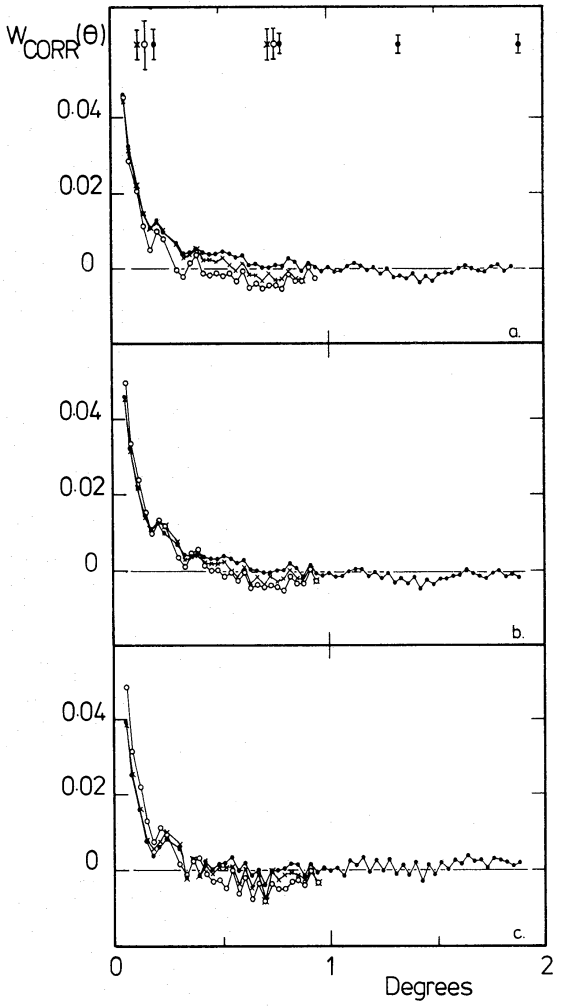

Figure 6. (a) Corrected estimators for the ensemble estimators of the DM 80 field shown in Fig. 3(a). The corrected curves are calculated according to equation (12) using the curves from Figs 3(a), 4(a) and 5(a). (b) Corrected estimators for the direct estimators of the DM 80 sample shown in Fig. 3(b) using the curves from Figs 3(b), 4(b) and 5(b). (c) Corrected estimators for the filtered estimators of the DM 80 sample shown in Fig. 3(c) using the curves from Figs 3(c), 4(c) and 5(c).

apparent anticlustering about clustered objects (e.g. Peebles 1980, sections 31 and 32). This explains why some samples show apparent anticlustering when strong clustering is present. However, the primary cause of the discrepancies evident in the data under discussion is the nature of the samples: the framework in which correlation functions are considered is one of homogeneous, isotropic and 'fair' samples and in practice none of these three conditions is generally met. Large-scale anisotropies are introduced into wide-angle galaxy samples – e.g. the Zwicky and Lick counts – and most catalogues suffer from various inhomogeneities due to selection effects, variable plate properties and human error. Intrinsic inhomogeneities may also be present on large scales – if the existence of structures $> 100H^{-1}$ Mpc postulated by Einasto, Joeveer & Saar (1980) is confirmed. Two examples of corrections due to severe anisotropies are those of Sharp (1979) to the Zwicky catalogue and that due to Seldner & Peebles (1981) in their recent analysis of the Cambridge 4C Survey. In this case an apparent anticorrelation due to instrumental effects is removed by a method not dissimilar in principle to that described in this paper.

The deeper narrow angle samples do not suffer so much from large-scale absorption

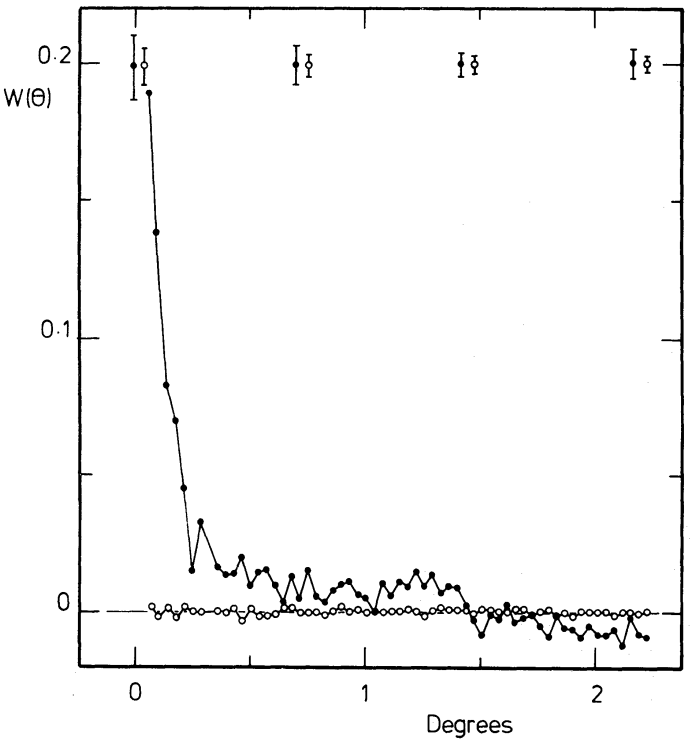

Figure 7. The ensemble estimator from the SGP V galaxy sample (•) together with the ensemble estimator from the parallel stellar sample (○) showing that the low-amplitude features in the galaxy covariance function are not due to any plate or machine measuring effects, unless the effects are only confined to galaxies.

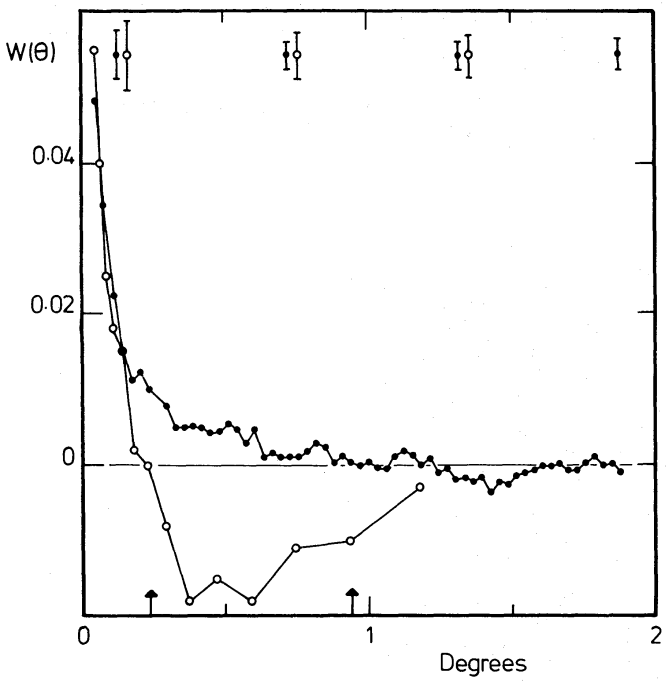

Figure 8. The ensemble estimator from the DM 80 sample (•) together with the original estimator from the DM 80 paper (○).

gradients but are particularly susceptible to a variety of telescopic, measuring machine, plate uniformity and possibly small-scale galactic absorption effects due to the strong dependence of galaxy surface density on limiting magnitude. Furthermore, although much deeper than the shallow samples, they cover only a relatively small volume at their typical depth of $Z^* \sim 0.2$–0.25, typically a sample from a Schmidt plate covers $40 \times 40 H^{-1}$ Mpc at $Z \sim 0.2$. Seldner & Peebles (1977) have shown that the haloes of rich clusters extend to $40 H^{-1}$ Mpc, and there is some evidence to indicate that very large superclusters may exist (e.g. Gregory *et al.* 1978; Tarenghi *et al.* 1980; Gregory, Thompson & Tifft 1981). Both of these considerations suggest that the samples analysed may generally not be fair and could be expected to show inhomogeneities and anisotropies on scales that would significantly affect the estimates of $W(\theta)$. The power of the ensemble estimators is that they remove large-scale gradients – whether due to the galaxy distribution or extrinsic causes – related to the sample boundaries, allowing examination of the properties of the sample which are unrelated to the boundaries. Consequently the ensemble estimators for the Jagellonian and SGP V samples discussed above will be unaffected by any large-scale gradients as these have been removed by the estimation procedure.

The confirmation of the existence of the break from the power-law form of the covariance function in both the Jagellonian and SGP V samples at scales in good agreement with that found by GP 77 is strong evidence for the reality of the feature. Also of interest is the detection of weak anticlustering at large scales in both the Jagellonian and SGP V samples. The corrected estimate of $W(\theta)$ for the Zwicky catalogue from Sharp (1979) appears to show some evidence of anticlustering. This is evident in fig. 1 of Bonometto & Sharp (1980), although it is not discussed. Examination of fig. 2 in Sharp (1979) suggests that $\hat{W}(\theta)$ becomes negative at a scale ~ 15 degrees. The onset of anticlustering is difficult to measure precisely because of the low amplitude of the effect, however, qualitative scaling of the feature is certainly present. The approximate scales in the Zwicky, Jagellonian and SGP V samples at which negative values of $\hat{W}(\theta)$ occur are 15 degrees, 1.5 degrees and 1.7 degrees corresponding to $13 H^{-1}$ Mpc, $10 H^{-1}$ Mpc and $11 H^{-1}$ Mpc respectively. The angular scales correspond to 19, 25 and 39 per cent of the sample diameter in each case. The onset of negative values of $\hat{W}(\theta)$ does not correspond to the same fraction of the sample size as might be expected if they were an artefact of the analysis technique.

Interpretation of the negative values of $\hat{W}(\theta)$ without further confirmation from larger samples is premature. The significance of the feature can not be assessed in any case until the full extent of the negative region is found, when a calculation similar to that of Peebles (1974) could be made. If the anticlustering were due to galaxies distributed in long filaments surrounded by large 'holes' (e.g. Gregory *et al.* 1981) then the three-point correlation function might show a strong dependence of the function on the parameters 'u' and 'v' – which describe the shape of the triangles of galaxies (see GP 77) – on scales of 15–$20 H^{-1}$ Mpc where the apparent anticlustering manifests itself in the two-point function. Larger galaxy samples are necessary to test this prediction. Of particular concern is the difference between estimates of $W(\theta)$ from the SGP V sample and the SFEM data of a similar field. The selection functions for both samples are well understood so the change in the linear scale of the break point from $8 H^{-1}$ Mpc (this paper) and $3 H^{-1}$ Mpc found by SFEM is disturbing. T. Shanks has kindly provided me with SFEM's deepest J sample of the SGP field in the form of a 32×32 array of cell counts. Analysis of this data broadly confirms the published result supporting the view that in the case of the SFEM study the discrepancies are a real feature of the data. More machine measured samples from deep plates are necessary before the discrepancies evident both between DM 80, MD 79 and SFEM and with earlier work can be regarded as firmly established.

7 Conclusions

The discrepant estimates of $W(\theta)$ for the DM 80 sample and almost certainly those of the MD 79 samples have been shown to be due to the analysis techniques employed. A new sample in the SGP suggests that the disagreement in power-law break length between the Princeton results and those of SFEM is due to some artefact of the COSMOS data or data reduction. The existence of the power-law break point at scales entirely consistent with that found by GP 77 in the Lick counts is established for the Jagellonian and a new SGP V sample. This contrasts with evidence from recent machine measures that the position of the break point and the general form of the covariance function at scales beyond $3H^{-1}$ Mpc are highly variable. A consequence of using ensemble estimators when calculating estimates of the covariance function – or correctly removing biases by cross-correlation with random data – is the detection of weak anticlustering in the Jagellonian and SGP V samples at angular scales corresponding to a linear size of order $15H^{-1}$ Mpc. The use of an ensemble estimation technique in conjunction with edge correction method 'A' of Section 3.2 to estimate $W(\theta)$ is strongly recommended for the calculation of all correlation functions. When an ensemble technique can not be applied directly it is essential that corrections according to the prescription of Section 3.3 are made. With the increasing use of correlation functions in many fields of astronomy it is important that the magnitude of systematic effects is realised and ensemble techniques employed to remove any bias.

Acknowledgments

I would particularly like to thank Dr D. Emerson for help and encouragement throughout the project. Professor M. Longair, Drs G. Gilmore, N. Sharp and T. Shanks made valuable comments on an earlier draft of this paper. Drs H. T. MacGillivray and R. Fong provided useful discussions. The UKSTU and IDPU units at the Royal Observatory Edinburgh together with Dr R. S. Ellis provided plates, measures and data. I acknowledge the receipt of a SERC studentship.

References

Bahcall, J. N. & Soneira, R. M., 1981. *Astrophys. J.*, **246**, 122.
Bonometto, S. A. & Lucchin, F., 1980. *Astr. Astrophys.*, **82**, 287.
Bonometto, S. A. & Sharp, N. A., 1980. *Astr. Astrophys.*, **92**, 222.
Dautcourt, G., Kempe, K., Richter, L. & Richter, N., 1978. *Astr. Nach.*, **299**, 177.
Davis, M., Groth, E. J. & Peebles, P. J. E., 1977. *Astrophys. J.*, **212**, L107.
Dodd, R. J. & MacGillivray, H. T., 1980. *ESO Workshop on Two-Dimensional Photometry*, p. 391, eds Craine, P. & Kjar, K., European Southern Observatory.
Einasto, J., Joeveer, M. & Saar, E., 1980. *Nature*, **283**, 47.
Ellis, R. S., Fong, R. & Phillipps, S., 1977. *Mon. Not. R. astr. Soc.*, **181**, 163.
Fry, J. N. & Peebles, P. J. E., 1978. *Astrophys. J.*, **221**, 19.
Gregory, S. A., Thompson, L. A. & Tifft, W. G., 1981. *Astrophys. J.*, **243**, 411.
Gregory, S. A., Chincarini, G., Rood, H. J. & Thompson, L. A., 1978. *Astrophys. J.*, **253**, 724.
Groth, E. J. & Peebles, P. J. E., 1977. *Astrophys. J.*, **217**, 385.
Longair, M. S. & Seldner, M., 1979. *Mon. Not. R. astr. Soc.*, **189**, 433.
MacGillivray, H. T. & Dodd, R. J., 1979. *Man. Not. R. astr. Soc.*, **186**, 69.
Mac Gillivray, H. T. & Dodd, R. J., 1980. *Mon. Not. R. astr. Soc.*, **193**, 1.
Masson, C., 1979. *Mon. Not. R. astr. Soc.*, **188**, 261.
Osmer, P., 1982. *Astrophys. J.*, in press.
Peebles, P. J. E., 1973. *Astrophys. J.*, **185**, 413.
Peebles, P. J. E., 1974. *Astr. Astrophys.*, **32**, 197.
Peebles, P. J. E., 1975. *Astrophys. J.*, **196**, 647.

Peebles, P. J. E., 1980. *The Large Scale Structure of the Universe,* Princeton University Press, Princeton.
Peebles, P. J. E. & Groth, E. J., 1975. *Astrophys. J.,* **196,** 1.
Peebles, P. J. E. & Hauser, M. G., 1974. *Astrophys. J. Suppl.,* **28,** 19.
Phillipps, S., Fong, R., Ellis, R. S., Fall, S. M. & MacGillivray, H. T., 1978. *Mon. Not. R. astr. Soc.,* **182,** 673.
Reid, I. R. & Gilmore, G., 1981. *Mon. Not. R. astr. Soc.,* **201,** 73.
Sargent, W. L. W., Young, P. J., Boksenberg, A. & Tytler, D., 1980. *Astrophys. J. Suppl.,* **42,** 41.
Seldner, M. & Peebles, P. J. E., 1977. *Astrophys. J.,* **215,** 703.
Seldner, M. & Peebles, P. J. E., 1981. *Mon. Not. R. astr. Soc.,* **194,** 251.
Seldner, M., Siebers, B., Groth, E. J. & Peebles, P. J. E., 1977. *Astr. J.,* **82,** 249.
Shanks, T., Fong, R., Ellis, R. S. & MacGillivray, H. T., 1980a. *Mon. Not. R. astr. Soc.,* **192,** 209.
Shanks, T., Phillipps, S. & Fong, R., 1980b. *Mon. Not. R. astr. Soc.,* **191,** 47P.
Sharp, N. A., 1979. *Astr. Astrophys.,* **74,** 312.
Tarenghi, M., Chincarini, G., Rood, H. J. & Thompson, L. A., 1980. *Astrophys. J.,* **253,** 724.