Research in Astron. Astrophys. **2011** Vol. **11** No. **6**, 655–670
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

On the fairness of the main galaxy sample of SDSS *

Ke-Lai Meng[1], Bin Ma[2], Jun Pan[1] and Long-Long Feng[1]

[1] Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008, China; *mkl@pmo.ac.cn*
[2] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Received 2010 September 15; accepted 2011 February 11

Abstract Flux-limited and volume-limited galaxy samples are constructed from the Sloan Digital Sky Survey (SDSS) data releases DR4, DR6 and DR7 for statistical analysis. The two-point correlation functions $\xi(s)$, monopole of three-point correlation functions ζ_0, projected two-point correlation function $w_{\rm p}$ and pairwise velocity dispersion σ_{12} are measured to test if galaxy samples are fair for these statistics. We find that with the increment of sky coverage of subsequent data releases in SDSS, $\xi(s)$ of the flux-limited sample is extremely robust and insensitive to local structures at low redshift. However, for volume-limited samples fainter than L^* at large scales $s \gtrsim 10\,h^{-1}\,{\rm Mpc}$, the deviation of $\xi(s)$ from different SDSS data releases (DR7, DR6 and DR4) increases with the increment of absolute magnitude. The case of $\zeta_0(s)$ is similar to that of $\xi(s)$. In the weakly nonlinear regime, there is no agreement between ζ_0 of different data releases in all luminosity bins. Furthermore, $w_{\rm p}$ of volume-limited samples of DR7 in luminosity bins fainter than $-M_{r,0.1} = [18.5, 19.5]$ are significantly larger and σ_{12} of the two faintest volume-limited samples of DR7 display a very different scale dependence than results from DR4 and DR6. Our findings call for caution in understanding clustering analysis results of SDSS faint galaxy samples and higher order statistics of SDSS volume-limited samples in the weakly nonlinear regime. The first zero-crossing points of $\xi(s)$ from volume-limited samples are also investigated and discussed.

Key words: galaxies: distances and redshifts — galaxies: statistics — cosmology: observation — cosmology: large-scale structure

1 INTRODUCTION

Clustering analysis of galaxy samples thrives on the availability of large modern galaxy surveys. The two largest and most successful galaxy surveys to date are the two-degree field galaxy redshift survey (2dFGRS, Colless et al. 2003) and the Sloan Digital Sky Survey (SDSS, York et al. 2000). The final data release of the 2dFGRS offers 3-D mapping of roughly a quarter of a million galaxies, while the SDSS has achieved spectra of ~ 0.9 million galaxies (Abazajian et al. 2009). The unprecedented number of galaxies and the enormous volume surveyed by SDSS defines its unique role in the era

* Supported by the National Natural Science Foundation of China.

of precision cosmology (Komatsu et al. 2011), especially considering its power spectra and the two-point correlation functions (2PCF) at large scales (e.g. Tegmark et al. 2004; Eisenstein et al. 2005; Percival et al. 2010; Reid et al. 2010).

Another highly appreciated application of clustering analysis of galaxies is to relate galaxy distribution to dark matter and halos, aiming at inferring processes galaxies experienced during their formation and evolution. Interpretation of statistics of galaxy samples provided by SDSS prevails in the category of the ΛCDM+halo model and relevant extensions such as the halo occupation distribution (HOD, e.g. Berlind & Weinberg 2002; Kravtsov et al. 2004; Zheng et al. 2005) and the conditional luminosity function (CLF, Yang et al. 2003). For example, works of Zehavi et al. (2002, 2005) and The SDSS Collaboration et al. (2010) systematically explored the luminosity and color dependence of galaxy 2PCFs and extensively quantified HOD parameters of galaxies; Cooray (2006) derived the occupation numbers of central and satellite galaxies in halos and their corresponding conditional luminosity functions from a compilation of correlation functions of SDSS, attempting to draw clues of galaxy evolution with reference to high redshift samples; Li et al. (2007) rather directly compared projected correlation functions and the pairwise velocity dispersion (PVD) from SDSS with those of mock galaxy samples populated from N-body simulations by semi-analytic models (SAM) of Kang et al. (2005) and Croton et al. (2006); they found that SAM can roughly reproduce observed clustering of SDSS galaxies but have to reduce the faint satellite fraction in massive halos using the prescription of SAM by ~ 30 percent to resolve discrepancies in PVD.

Yet there are challenges to the fairness of SDSS galaxy samples, i.e. whether galaxy samples of SDSS are complete and have enough volume to be a fair representation of the Universe. In fact, prudence in extracting physics from measured statistics, especially correlation functions, has been called upon. Nichol et al. (2006) disclosed that exclusion of the *Sloan Great Wall* (at $z \sim 0.08$, Gott et al. 2005) would change the 2PCF by $\sim 40\%$ and the three-point correlation function (3PCF) by as much as $\sim 70\%$ for the sample defined by the *r*-band absolute magnitudes $-22 \leq M_{r,0.1} \leq -19$. The apparent influence of super structures on estimated correlation functions at large scales somehow counters intuition since one already takes it for granted that the SDSS galaxy sample's depth and sky coverage are sufficient to accomplish homogeneity, and spatial averaging would suppress the variance induced by a particular structure in a small patch of sky. Sylos Labini et al. (2009) noticed that the zero-crossing point of 2PCF in the SDSS main galaxy sample varies with luminosity and sample depth and anti-correlation is absent in the most recently measured 2PCF of the SDSS luminous red galaxy (LRG) sample (e.g. Martínez et al. 2009; Kazin et al. 2010). From implementing methods of extreme-value statistical analysis, Antal et al. (2009) purport that either the SDSS suffers from severe sample volume dependent intrinsic systematical effects or there is persistent density fluctuation not fading away over scales beyond the standard ΛCDM model prediction.

It is therefore important for one to check the fairness of galaxy samples used in order to endorse the confidence of relevant analysis. It is understood that fairness means different results for different statistical method and different samples constructing methods. The SDSS Collaboration et al. (2010) laboriously evaluated finite volume effects and impacts of super structures, and they compared 2PCFs of volume-limited galaxy sub-samples in full depth with the same sub-sample but limited to a smaller volume overlapping with the volume-limited sub-sample defined in the luminosity bin one dex lower. Their experiment leads to the conclusion that finite volume effects are insignificant for anisotropic and projected 2PCFs in the nonlinear regime of their sub-samples of luminosity higher than $M_{r,0.1} = -19$. However, they then found that including faint galaxies causes weird behavior of the 2PCF, which is similar to the discovery in Zehavi et al. (2005) using an early release of SDSS, but with a smaller amplitude. We notice that such analyses for galaxies with luminosity lower than -18 are missed though 2PCF of their faintest sub-sample $M_{r,0.1} \in [-18, -17]$ is adopted for estimation of biasing and HOD parameters.

These works mainly concentrate on changes to two-point statistics by altering sample depth. We prefer to check the fairness by sky coverage enlargement, not only of 2PCFs but also of the

monopole in 3PCFs in redshift space, projected 2PCFs and PVDs. There are data releases 4, 6 and 7 of SDSS' main galaxy catalog (DR4, DR6 and DR7 by Adelman-McCarthy et al. 2006, 2008; Abazajian et al. 2009, respectively), with the increment of sky coverage from DR4 to DR6 being roughly the same from DR6 to DR7. An advantage of investigating effects of sample volume on the correlation function with sky angular coverage, excluding the effects of survey depth, is that the restriction of apparent magnitudes in the survey scope limits the permitted range of depth. This is especially true for those galaxies which are visible only at low redshift or span a very shallow range in the sample space. One of our purposes is to see how the correlation function evolves *naturally* with the development of a real survey.

Section 2 describes SDSS data and estimation methods of statistics we used. Results are shown in Section 3. The last section is for summary and discussion.

2 GALAXY SAMPLES AND ESTIMATION OF CORRELATION FUNCTIONS

2.1 Sample Construction

The `safe` galaxy sample of the New York University Value-Added Galaxy Catalog (NYU-VAGC, Blanton et al. 2005) [1] is a catalog of low redshift galaxies (mostly below $z \sim 0.3$) defined by apparent magnitudes of $14.5 < m_r < 17.6$. Three data releases in chronological order are selected, namely DR4, DR6 and DR7, which spectroscopically surveyed areas of about $4\,783$, $6\,860$ and $8\,032$ square degrees, respectively. Since spectroscopic coverage of SDSS is not uniform, we use only those regions of spectroscopic completeness greater than 0.9. We did not perform fiber collision correction to improve completeness, since the correction only becomes significant at scales $< 0.2\,h^{-1}$ Mpc for SDSS galaxies (Zehavi et al. 2002). To ensure the correct geometry, galaxies in the three catalogs are also filtered with their own accompanying survey windows, bright star masks and completeness masks.

Flux-limited samples defined by the *r*-band apparent magnitude range $14.5 < m_r < 17.6$ and redshift $0.01 < z < 0.23$ are generated. Consequently, we obtain $300\,661$ galaxies in DR4, $447\,407$ in DR6, and $535\,845$ in DR7. In order to explore the influence of local galaxies on correlation functions, we also constructed flux-limited galaxy samples by near-end redshift cuts of $z_{\rm min} = 0.037, 0.046$ and 0.071. Volume-limited sub-samples are also produced in consecutive luminosity bins starting from $M_{r,0.1} = -17$ to -22.5 in steps of 0.5 magnitude and bin widths of one magnitude. The absolute magnitude in NYU-VAGC is corrected to redshift $z = 0.1$ and is K corrected, but e-correction is not taken into account. We noticed that there are some galaxies that have different apparent magnitudes in DR7 than in earlier data releases, so we constructed a couple of additional volume-limited samples from DR7 but filtered the samples from DR4 with masks for comparison. Measurements indicate that such differences have little influence on the statistics employed.

Table 1 Numbers of galaxies in flux limited samples defined by the *r*-band apparent magnitude range $14.5 < m_r < 17.6$ and redshift range $z_{\rm min} \le z \le 0.23$.

$z_{\rm min}$	0.010	0.037	0.046	0.071
DR4	300 661	281 400	268 247	216 373
DR6	447 407	417 426	397 543	321 915
DR7	535 845	498 445	473 980	382 921

Details of these samples are shown in Tables 1 and 2 and Figure 1; comoving distances of galaxies are calculated in a flat ΛCDM universe with $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$ and $h = 0.7$.

[1] *http://sdss.physics.nyu.edu/vagc*

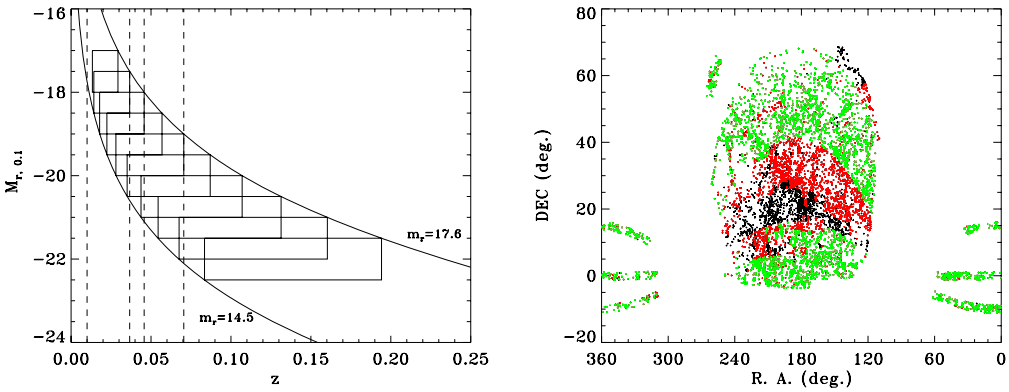

Fig. 1 *Left* panel shows the definition of SDSS galaxy subsamples in the redshift-absolute magnitude plane. The two curves are boundaries of the VAGC catalog resulting from the imposed apparent magnitude limits. The overlapping rectangles delineate where volume-limited samples are located and dashed lines label the lower redshift cuts of our flux-limited samples. In the *right* panel, distributions of galaxies of the volume-limited subsample $-M_{r,0.1} = [17, 18]$ on the celestial sphere are plotted. *Green* points are galaxies from SDSS DR4, *red* points indicate extra galaxies in DR6 and black points are galaxies added in DR7 (color online).

Table 2 Volume limited samples. Distances are in units of h^{-1} Mpc.

Label	Luminosity	Redshift		Comoving distance		Number of galaxies		
	$M_{r,0.1} - 5\log_{10} h$	$z_{\min}$	$z_{\max}$	$d_{\min}$	$d_{\max}$	DR4	DR6	DR7
VL1	$[-18.0, -17.0]$	0.011	0.029	33.89	87.31	4 223	6 389	8 219
VL1+	$[-18.5, -17.5]$	0.014	0.037	42.53	108.95	7 292	11 543	14 343
VL2	$[-19.0, -18.0]$	0.018	0.046	53.32	135.61	11 639	18 328	22 500
VL2+	$[-19.5, -18.5]$	0.022	0.057	66.77	168.27	19 209	29 463	35 932
VL3	$[-20.0, -19.0]$	0.028	0.071	83.51	207.96	31 807	47 565	57 363
VL3+	$[-20.5, -19.5]$	0.035	0.087	104.24	255.70	50 719	75 162	89 654
VL4	$[-21.0, -20.0]$	0.044	0.107	129.83	312.96	59 215	87 295	103 924
VL4+	$[-21.5, -20.5]$	0.054	0.131	161.21	381.91	60 132	89 602	107 207
VL5	$[-22.0, -21.0]$	0.068	0.160	199.41	462.71	46 264	69 499	82 239
VL5+	$[-22.5, -21.5]$	0.083	0.194	245.46	555.88	24 002	36 677	43 631

2.2 Estimation of Correlation Functions

2.2.1 Redshift space correlation functions

Isotropic 2PCF $\xi(s)$ of separation s in redshift space is measured with the estimator of Landy & Szalay (1993),

$$\xi = \frac{\mathrm{DD} - 2\mathrm{DR} + \mathrm{RR}}{\mathrm{RR}}\,, \tag{1}$$

in which DD, RR and DR are respectively the normalized numbers of weighted galaxy-galaxy, random-random and galaxy-random pairs at given separations. To proceed with the estimation using Equation (1), the corresponding random sample is generated following distributions of redshift, magnitude, geometric constraints, spectroscopic completeness and survey masks of each individual galaxy sample but with twenty times the number of points. Each galaxy and random point is assigned a weight according to their redshift and angular position to minimize the variance in estimating ξ

(Efstathiou 1988; Hamilton 1993),

$$w_i = \frac{1}{1 + 4\pi\, n(z)\Phi_i\, J_3(s)}\ , \tag{2}$$

where Φ_i is the selection function at the location of the ith galaxy, $n(z)$ is the mean number density and $J_3(s) = \int_0^s \xi(s)s^2 ds$. The $J_3(s)$ is computed using a power-law $\xi(s)$ with correlation length $s_0 = 8\,h^{-1}$ Mpc and $\gamma_0 = 1.2$ (Zehavi et al. 2002).

The calculation of 3PCFs of all those galaxy samples takes too long, so we measured the monopole of the 3PCF instead (Pan & Szapudi 2005), which is a degenerate version of 3PCF defined as

$$\zeta_0(s_1, s_2) = 2\pi \int_{-1}^{1} \zeta(s_1, s_2, \theta) \mathrm{d}\cos\theta\ , \tag{3}$$

and estimated via

$$\zeta_0 = \frac{\mathrm{DDD} - 3\mathrm{DDR} + 3\mathrm{DRR} - \mathrm{RRR}}{\mathrm{RRR}}\ , \tag{4}$$

where combined symbols of D and R are normalized numbers of triplets counted within and between data sets of galaxies and random points, e.g. if the number of galaxies around galaxy i in bin $(s_1^{\mathrm{lo}}, s_1^{\mathrm{hi}})$ is $n_i(s_1)$, and the number in bin $(s_2^{\mathrm{lo}}, s_2^{\mathrm{hi}})$ is $n_i(s_2)$, the DDD in Equation (4) reads

$$\mathrm{DDD} = \begin{cases} \dfrac{\sum_{i=1}^{N_\mathrm{g}} n_i(s_1) n_i(s_2)}{N_\mathrm{g}(N_\mathrm{g}-1)(N_\mathrm{g}-2)} & \text{if } \ s_1 \neq s_2\,, \\ \dfrac{\sum_{i=1}^{N_\mathrm{g}} n_i(s_1)\,(n_i(s_2)-1)}{N_\mathrm{g}(N_\mathrm{g}-1)(N_\mathrm{g}-2)} & \text{if } \ s_1 = s_2\,. \end{cases} \tag{5}$$

2.2.2 Projected 2PCF and PVD

To minimize the effect of redshift distortion due to a galaxy's peculiar motion, the separation s (or r in real space) is divided into two components: the parallel part π and the perpendicular part σ with respect to line-of-sight. The anisotropic 2PCF is measured on grids of (σ, π). Integration of $\xi(\sigma, \pi)$ over π then yields a distortion-free redshift function, the projected 2PCF,

$$w_\mathrm{p}(\sigma) = \int_{-\pi_\mathrm{max}}^{+\pi_\mathrm{max}} \xi(\sigma, \pi) d\pi = \sum_i \xi(\sigma, \pi_i)\Delta\pi_i\ , \tag{6}$$

which practically has an integration limit of $\pi_\mathrm{max} = 50\,h^{-1}$ Mpc.

It is well known that the redshift distortion consists of two components which dominate in different regimes. Coherent infall is responsible for the clustering enhancement at large scales while the smearing of correlation strength at small scales is attributed to random motions. At large scales, the boost to the 2PCF by the peculiar velocities takes a particularly simple form (Kaiser 1987; Hamilton 1992),

$$\xi'(\sigma, \pi) = \xi_0(s)P_0(\mu) + \xi_2(s)P_2(\mu) + \xi_4(s)P_4(\mu)\ , \tag{7}$$

where $P_\ell(\mu)$ represents Legendre polynomials, and μ is the cosine of the angle between r and π. Assuming $\xi = (r/r_0)^{-\gamma}$, there are relations

$$\begin{aligned} \xi_0(s) &= \xi(s) = \left(1 + \frac{2\beta}{3} + \frac{\beta^2}{5}\right)\xi(r)\,, \\ \xi_2(s) &= \left(\frac{4\beta}{3} + \frac{4\beta^2}{7}\right)\left(\frac{\gamma}{\gamma-3}\right)\xi(r)\,, \\ \xi_4(s) &= \frac{8\beta^2}{35}\left(\frac{\gamma(2+\gamma)}{(3-\gamma)(5-\gamma)}\right)\xi(r)\ , \end{aligned} \tag{8}$$

where $\beta \approx \Omega_0^{0.6}/b$ and b is the linear bias parameter; note that the first equation is independent of the functional form of $\xi(r)$.

To incorporate effects of random motion, the anisotropic 2PCF in redshift space is approximated by a convolution of $\xi'(\sigma,\pi)$ in Equation (7) with the distribution function of the pairwise velocity $f(v_{12})$ (c.f. Peebles 1993),

$$\xi(\sigma,\pi) = \int_{-\infty}^{+\infty} \xi'(\sigma,\pi - \frac{v_{12}}{H_0}) f(v_{12}) dv_{12} \ , \tag{9}$$

and in general $f(v_{12})$ is assumed to obey an exponential distribution with PVD σ_{12}

$$f(v_{12}) = \frac{1}{\sigma_{12}\sqrt{2}} \exp\left(-\frac{\sqrt{2}v_{12}}{\sigma_{12}}\right) \ . \tag{10}$$

The parameter β is usually derived from the ratio of $\xi(s)$ to $\xi(r)$ at large scales via the first equation in Equation (8), then other model parameters can be determined by combining Equations (7) – (10) to fit the $\xi(\sigma,\pi)$ data grids. Note that Jing et al. (1998) assumed a slightly different exponential distribution function for pairwise velocity which was followed by Li et al. (2007).

2.2.3 Covariance matrix

Covariance matrices of our results are computed with the jackknife technique (Lupton 1993; Zehavi et al. 2002). Each galaxy sample is divided into twenty separate slices of approximately equal sky area, then we perform the analysis twenty times, leaving a different slice out each time. Covariance matrices are generated accordingly with these twenty measurements, for instance, the covariance of 2PCF measured in two bins i and j is simply

$$\mathrm{Cov}(\xi_i,\xi_j) = \frac{N-1}{N} \sum_{\ell=1}^{N} (\xi_{i,\ell} - \overline{\xi_i})(\xi_{j,\ell} - \overline{\xi_j}) \ , \tag{11}$$

in which $N = 20$ is the number of jackknife sub-samples we used.

3 RESULTS

3.1 Flux-limited Samples

Isotropic 2PCFs of flux-limited samples in Table 1 are calculated first. Figure 2 demonstrates that the redshift space of 2PCFs for flux-limited samples shows little variation between various data versions of SDSS. $\xi(s)$ of DR4 exhibits some deviation at large scales $\sim 100\,h^{-1}\,\mathrm{Mpc}$, but is hardly significant for the huge cosmic variance at these scales. $\xi(s)$ of flux-limited samples of the same data release are displayed in the right panel of Figure 2, and there is no visible change to redshift space 2PCF of SDSS when galaxies with low redshift are excluded, even when the decrease in the number of galaxies is as much as $\sim 25\%$ (Table 1). Since eliminating local volume and enlarging sky coverage from DR4 to DR7 have little influence on the clustering strength measured, it is unlikely that there are any significant sample volume dependent effects. Because we are not interested in a general discussion of the SDSS main galaxy catalog as a whole, we will stop performing further analysis with other statistical measures.

It is well known that faint galaxies have a much lower linear bias than luminous ones (e.g. Tegmark et al. 2004; Zehavi et al. 2005; Li et al. 2007). When we discard many faint galaxies by imposing a near-end redshift limit, it is expected that $\xi(s)$ should display higher amplitude when lower redshift cuts increase. It could be that the loss in number of galaxies (after proper weighting) is too small to raise any serious deviation (Table 1), or in other words $\xi(s)$ of the flux-limited sample is dominated by galaxies around the redshift distribution peak.

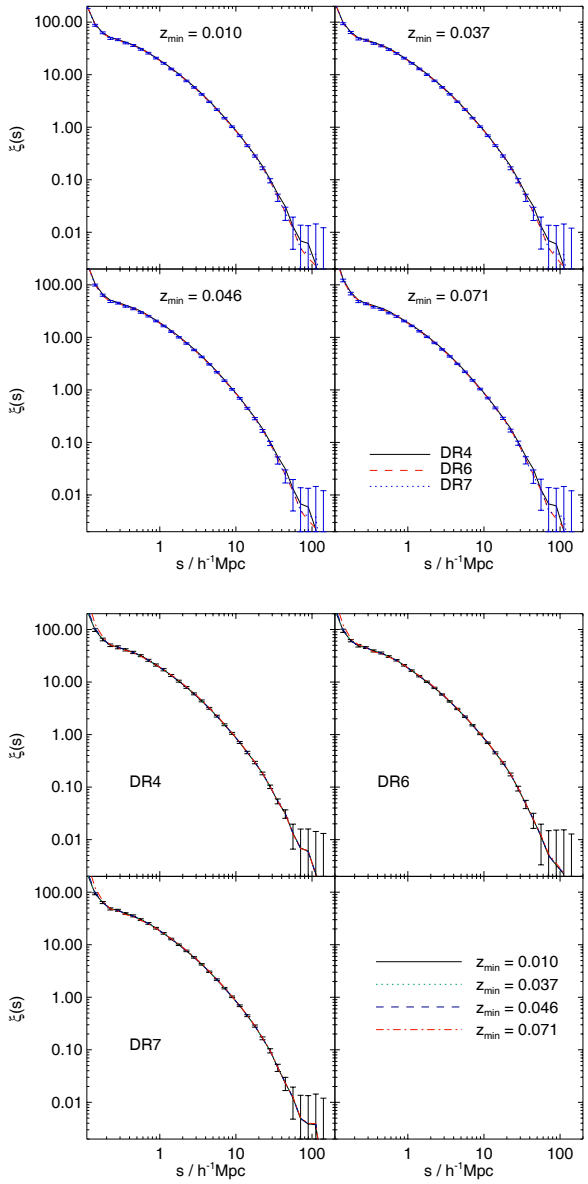

Fig. 2 Redshift space 2PCFs of flux-limited samples.

3.2 Volume-limited Samples

3.2.1 2PCF and the monopole of 3PCF in redshift space

The 2PCFs $\xi(s)$ and monopoles of 3PCFs $\zeta_0(s_1, s_2)$ from volume-limited samples of the three SDSS data releases are measured to probe possible differences. In this paper, we only present the $\zeta_0(s_1 = s_2)$ whose amplitude is the strongest among configurations of (s_1, s_2) (Pan & Szapudi 2005). As seen in Figure 3, in the nonlinear regime major discrepancies appear in the VL1 sample of the lowest luminosity; differences between results of DR4 and DR7 are around 2σ at scales as small

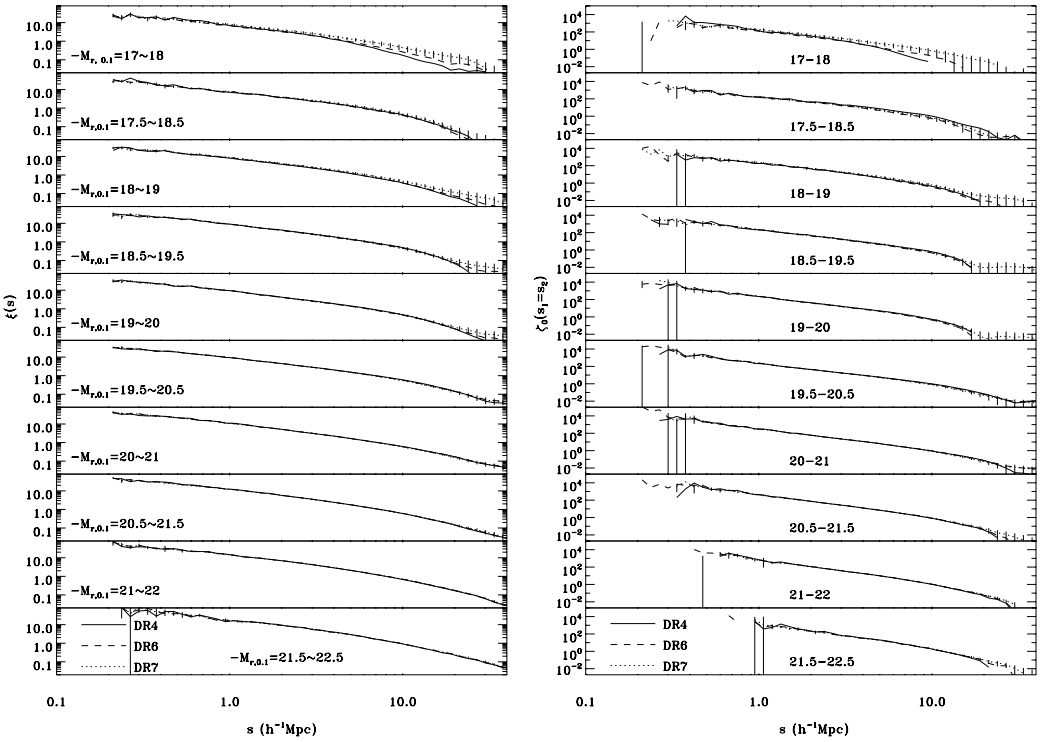

Fig. 3 2PCFs $\xi(s)$ and monopoles of 3PCF $\zeta_0(s_1 = s_2)$ at small scales in redshift space of volume-limited samples.

as $\sim 3h^{-1}$ Mpc, while the consistency of $\xi(s)$ and ζ_0 of brighter volume-limited samples of SDSS is perfect at scales $s < 10\,h^{-1}\,\mathrm{Mpc}$.

At scales greater than $10\,h^{-1}\,\mathrm{Mpc}$, for subsamples of VL3+ – VL5+, $\xi(s)$ values from different data releases are in good agreement within error bars, but ζ_0 values have variations at the level of $\sim 1\sigma$ (Fig. 4). For the five faint galaxy samples of VL1 – VL3, disagreement in $\xi(s)$ of DR7 to DR4 is already apparent in this regime, which is confirmed by their ζ_0. We conclude that modulation of correlation functions in redshift space resulting from enlargement of sky coverage mainly occurs at scales ranging roughly from ~ 10 to $\sim 50\,h^{-1}\,\mathrm{Mpc}$, which is usually classified as the weakly nonlinear regime in structure formation theory. Those applications and their associated conclusion appear rather suspicious based on 3PCFs of volume-limited samples of SDSS at large scales. For three-point correlation functions in redshift space, fairness of volume-limited samples is guaranteed only at small scales, i.e. in the strongly nonlinear regime.

3.2.2 The first zero-crossing points of 2PCFs

To investigate the charge of Sylos Labini et al. (2009), the first zero crossing scales of $\xi(s)$ with respect to median luminosity of volume-limited samples are plotted in Figure 5. Estimated $\xi(s)$ is effectively averaged over a scale bin $[s^{\mathrm{lo}}, s^{\mathrm{hi}}]$ and the quoted scale is set to be $s = \sqrt{s^{\mathrm{lo}} s^{\mathrm{hi}}}$. It is unlikely that we could correctly find all zero points of $\xi(s)$ with our scale binning, so we choose to

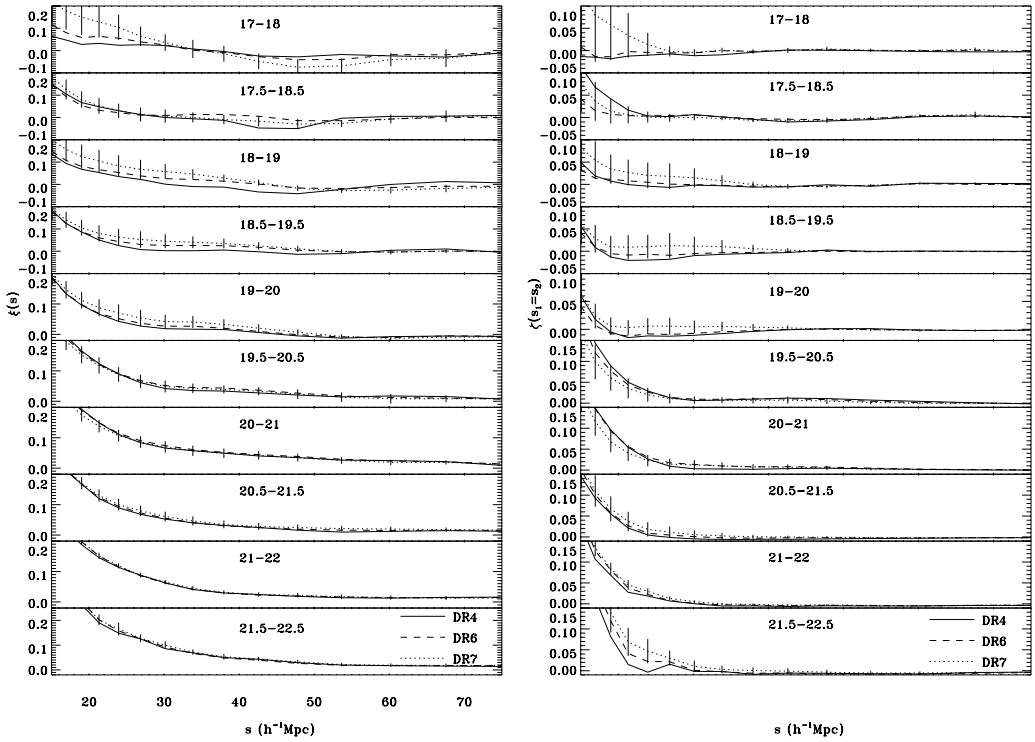

Fig. 4 2PCFs $\xi(s)$ and monopoles of 3PCFs $\zeta_0(s_1 = s_2)$ at large scales in redshift space of volume-limited samples.

show the range of scales within which $\xi(s)$ experiences zero-crossing, which is drawn as error bars over the geometric mean of the pair of scales. From Figure 5, it is clear that in general the brighter the characteristic luminosity of the sample is, the larger the first zero crossing scale will be. The five faint volume-limited samples (VL1 – VL3) have roughly the same first zero crossing scale with mild variation between $\sim 30 - 50\,h^{-1}\,\mathrm{Mpc}$, then the crossing scale ascends abruptly to as large as more than $100\,h^{-1}\,\mathrm{Mpc}$, even higher than the largest scale we measured ($\sim 170\,h^{-1}\,\mathrm{Mpc}$).

For faint galaxy samples, their depths are typically small and so are their effective volumes, so the systematical effect of integral constraint cannot be ignored (Landy & Szalay 1993; Bernstein 1994). In the weak correlation limit, the cosmic bias resulting from the integral constraint can be approximated by

$$b_\xi = \frac{\hat{\xi}}{\xi} - 1 \approx -\frac{\bar{\xi}(R)}{\xi}, \qquad \text{if} \quad |\xi|,\ |\bar{\xi}(R)| \ll 1\ , \tag{12}$$

in which $\hat{\xi}$ is the estimated 2PCF, R is the smallest size of the sample and $\bar{\xi}(R)$ is the average of the 2PCF over the sample volume, i.e. density variance at the sample volume (Landy & Szalay 1993). There is no *a priori* correction method to this bias unless we assume something to model the shape of the 2PCF. Since $\bar{\xi}$ is positive, naturally $\hat{\xi} \approx \xi - \bar{\xi}(R)$ will have a smaller first zero-crossing scale than ξ. If as usual we assume that galaxy bias b is linear and scale independent, $\hat{\xi} = b^2(\xi - \bar{\xi}(R))$, the correction to the first zero-crossing scale only depends on the sample volume. As $\bar{\xi}$ slowly decreases with scale, it is expected that the first zero-crossing scales of faint galaxy

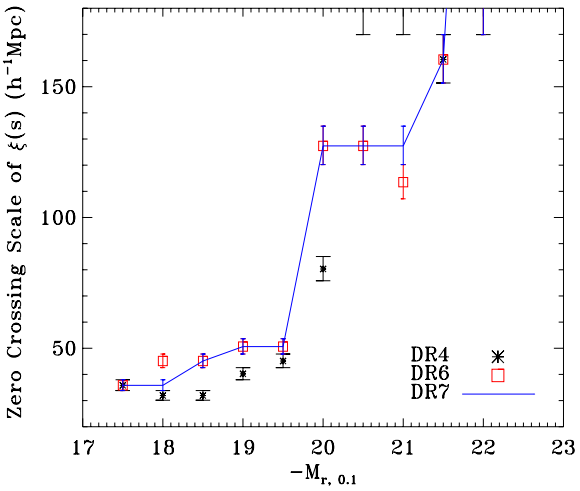

Fig. 5 Luminosity dependence of the first zero crossing scales of $\xi(s)$ of volume-limited samples. Lower caps of error bars are scales where $\xi > 0$ and higher caps of error bars are the adjacent scales where ξ immediately becomes negative. Those points shown with only lower caps denote that the first zero crossing point is actually larger than the scale probed in this work, larger than $\sim 170\,h^{-1}$ Mpc.

samples will gradually become larger when the sample volume increases, which is true for VL2 – VL3. However, surprisingly, this is not the case for VL1 and VL1+. The two faintest subsamples have the smallest sample volume, but the first-zero crossing scale of VL1 does not change from DR4 to DR7, but for VL+, the scale of DR7 becomes smaller than that of DR4. Furthermore, the difference between depths of VL3 and VL3+ is not very large (Table 2), but the first zero-crossing scales of their ξ differ greatly. The integral constraint alone could not explain these findings.

The increase in sky coverage from DR4 to DR6 is approximately the same as the gain from DR6 to DR7. The first zero crossing scales of DR7 differ only slightly from DR6 in two luminosity bins, but DR4 significantly disagrees with other data releases, which makes it difficult to believe another simple geometric explanation, such as assuming a fractal galaxy distribution. Ergodicity bias could not be used as an explanation. For low luminosity samples with low characteristic redshift, the correction $\Delta\xi$ is positive (Pan & Zhang 2010) and would push the zero point to larger scales, which obviously contradicts observation. Redshift distortion could also not be used, since on large scales redshift distortion acts on galaxy 2PCF as a multiplication factor.

The sudden change of the first zero-crossing scale from faint galaxies to bright galaxies probably implies that the composition of faint galaxy samples is very different compared with bright galaxy samples, which may be attributed to shifting the leading role from satellite galaxies to central galaxies in samples brighter than $-M_{r,0.1} > 20$ (Li et al. 2007). Whatever the physical mechanism is, mathematically the effect on 2PCF is encapsulated into a simple function: the galaxy bias. The linear biasing model assumes that on large scales the galaxy 2PCF with $\xi_{\rm g} = b\xi_{\rm m}$ in which $b \neq 0$ is a deterministic, scale independent bias parameter and $\xi_{\rm m}$ is the 2PCF of dark matter; obviously if the model holds, the zero point of $\xi_{\rm g}$ will not change no matter what b could be, e.g. scale dependent. If we presume that the problem of zero crossing is in biasing, then either stochastic or nonlinear bias has to be invoked. Simple calculation indicates that if we adopt the parametrization to bias the re-

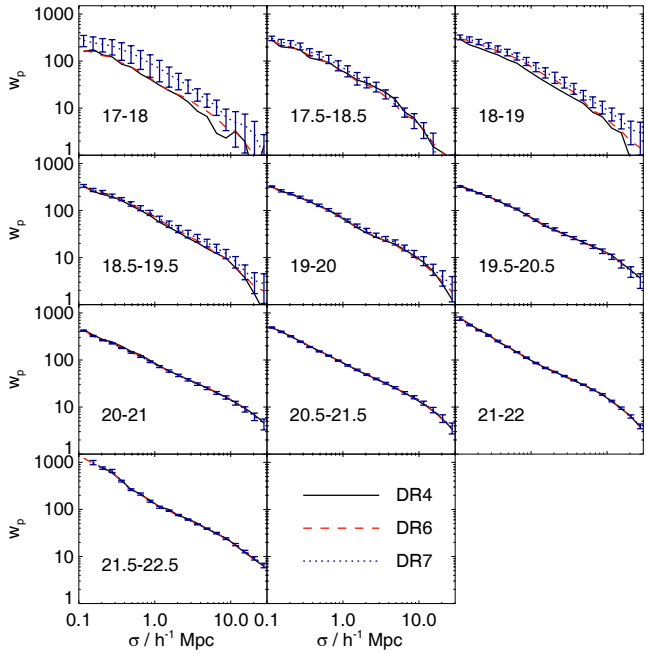

Fig. 6 Projected 2PCFs of the volume-limited samples.

sults of (Fry & Gaztanaga 1993) and include the second-order bias parameter in 2PCF, to the leading order the effect is again multiplicative and it cannot shift the first zero point of 2PCF. It appears that stochastic biasing has to be considered. Details of the calculation are, however, beyond the scope of this paper and will be presented elsewhere.

Another interesting aspect is that the first zero-crossing scale 2PCFs of samples VL4 and VL4+ of DR4 are larger than the largest scale of our measurements, but not of DR6 and DR7. The lack of anti-correlation in the two luminosity bins of DR4 is probably evidence of the modulation due to the Sloan Great Wall as revealed by Zehavi et al. (2005) and Nichol et al. (2006). The increased sky coverage of DR6 and DR7 just weakens the influence of the super structure (The SDSS Collaboration et al. 2010).

3.2.3 Projected 2PCF and PVD

$\xi(s)$ is a mixture of real space 2PCF and PVD. The entanglement can be sorted with the projected 2PCF $w_{\rm p}$. Measurements of $w_{\rm p}$ are shown in Figure 6. In fact, we cross checked our $w_{\rm p}$ of DR7 with available results from The SDSS Collaboration et al. (2010), and the agreement is excellent, except for the sample VL2 for which our $w_{\rm p}$ differs at scales $\sigma \gtrsim 4\,h^{-1}$ Mpc. As seen in Figure 6, it is obvious that $w_{\rm p}$ from DR4 and DR6 are in good agreement at the scale range probed in the most luminous bins; $w_{\rm p}$ values from DR6 are slightly larger at large scales around $\sigma \sim 10\,h^{-1}$ Mpc in several faint samples but have low significance because of the size of error bars. For VL1 and VL2, their $w_{\rm p}$ values from DR7 are boosted by more than 70% in amplitude relative to DR4, but the shape does not change. For subsamples in other luminosity bins, their $w_{\rm p}$ values are stable against data version, though for VL1+ and VL2+ there are some minor changes within error bars.

Figure 7 demonstrates the scale dependence of PVDs σ_{12} for different luminosity samples while Figure 8 has the luminosity dependence of PVDs measured at scales of $\sigma = 0.27, 0.87, 2.7$, and

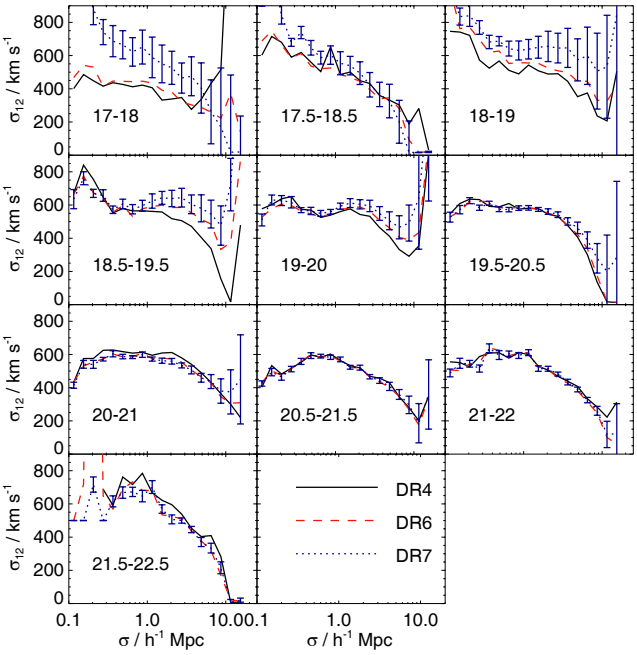

Fig. 7 Scale dependence of pairwise velocity dispersions in the volume-limited samples.

$8.7\,h^{-1}\,\mathrm{Mpc}$, respectively. Also, σ_{12} of subsamples VL1, VL2 and VL2+ from DR7 are significantly different compared to measurements of DR4. For VL2+, PVD from DR7 agrees with earlier data at small scales but then turns out to be higher at scales $\sigma > 1h^{-1}$ Mpc, which makes the scale dependence very weak; for VL2, σ_{12} of DR7 roughly keeps the shape of DR4 but has a much larger amplitude. Also, σ_{12} of VL1 from DR7 has a steeper scale dependence and stronger amplitude at small scales than results using DR4 and DR6. In addition, σ_{12} of VL1 subsamples from DR4 and DR6 are rather flat and do not follow the general trend that PVDs of galaxy samples with lower luminosities should rise faster at smaller scales (also see PVDs of SAMs in fig. 5 of Li et al. 2007), but now the VL1 result from DR7 reverts to this general trend that PVDs of galaxy samples with lower luminosities should rise faster at smaller scales. Comparing distributions in the celestial sphere of galaxies for the lowest luminosity bin of the three SDSS data releases reveals the variation is just induced by a large structure located roughly in an area of RA $166^\circ - 188^\circ$ and DEC of $16^\circ - 26^\circ$ (Fig. 1). This represents another example of the impact of super structure on clustering analysis of Large Scale Structure in addition to the Sloan Great Wall.

Li et al. (2007) realized that w_p and PVDs of faint volume-limited samples of DR4 are too low to match the prediction of SAMs. Guided by the experiment of Slosar et al. (2006), Li et al. (2007) reduced the fraction of satellite galaxies in massive halos in SAMs, in an *ad hoc* manner, by around 30% and approximately reproduced the actual measurements, which then became a serious challenge for researchers to reconcile disparities between models and observation. An eyeball check of our results with the SAMs prediction in Li et al. (2006) demonstrates that the amplitude boost in w_p and PVDs of DR7's faint volume-limited samples roughly compensate for the space between DR4 and SAMs, or at least ameliorate difficulties in theoretical modeling, although we do not have data on SAMs to quantify the improvement. So unlike the Sloan Great Wall, the existence of a large structure in the Universe is actually positive for our working models, which somehow casts doubts on the proclaimed practice of cutting off super structures from original data to better fit a unified

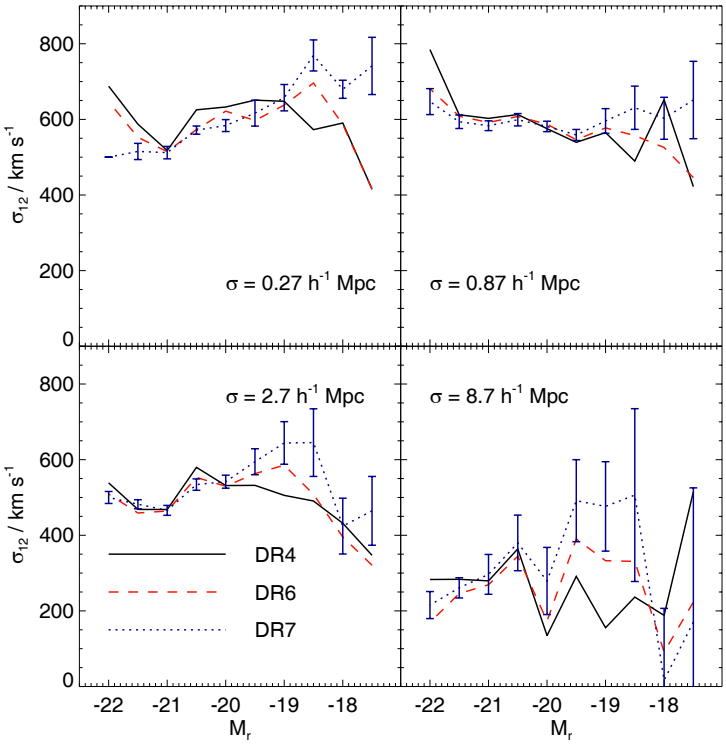

Fig. 8 Luminosity dependence of pairwise velocity dispersions at fixed scales.

picture. After all, it is still too early to say which is closer to the true clustering property of those faint galaxies, and we might need a deeper and wider survey than the present SDSS DR7 to reach reasonable fairness and reduce huge uncertainties.

4 SUMMARY AND DISCUSSION

By extensive comparison of different data releases of the SDSS main galaxy catalog with 2PCFs in redshift space for flux-limited samples, 2PCFs/monopoles of 3PCFs in redshift space for volume-limited samples, and projected 2PCFs and PVDs for volume-limited samples, we have the following findings about galaxy clustering properties with respect to the expansion of sky coverage of SDSS.

(1) 2PCFs $\xi(s)$ in redshift space of the flux-limited sample is extremely robust against sample volume change, which subsequently enables relevant application; $\xi(s)$ is also insensitive to local structures at low redshift.

(2) 2PCFs $\xi(s)$ in redshift space of volume-limited samples of SDSS DR7 in luminosity bins brighter than $-M_{r,0.1} = [17, 18]$ are in good agreement with earlier data releases at scales $s \lesssim 10\,h^{-1}$ Mpc. As scales become larger, the consistency is broken for volume-limited samples fainter than $-M_{r,0.1} = [19.5, 20.5]$ and in general the deviation of DR7 compared to DR6 and DR4 grows with larger absolute magnitude. Zero crossing points of DR7's $\xi(s)$ do not differ much compared to DR6's values, but apparently shift away from DR4's ones.

(3) Volume-limited samples of SDSS display convergence in ζ_0 at scales $s \lesssim 10\,h^{-1}$ Mpc, except the one in the faintest luminosity bin, but in the weakly nonlinear regime, there is no agreement between ζ_0 from different data releases in all luminosity bins.

(4) Projected 2PCFs' $w_{\rm p}$ of volume-limited samples in luminosity bins brighter than $-M_{r,0.1} = [18.5, 19.5]$ are robust with respect to data version, but for samples in fainter bins, $w_{\rm p}$ of DR7 are significantly higher than those of earlier data. A similar phenomenon is also seen in PVDs. PVDs of the two faintest volume-limited samples also appear much steeper along the scale in DR7 and then become flatter at higher luminosity, which actually turns out to be closer to what SAMs predict, as shown in Li et al. (2007).
(5) The faintest volume-limited sample of $-M_{r,0.1} = [17, 18]$ is very peculiar. It suffers from the biggest variance due to enlargement of sky coverage. The $\xi(s)$ and $\zeta_0(s)$ from DR7 agree well with the results from early data at large scales, but at scales as small as $\sim 3\,h^{-1}$ Mpc the agreement stops; $w_{\rm p}$ of the sample is enhanced around $\sim 70\%$. PVDs are rather distinguished in amplitude and scale dependence from measurements of earlier data.

Fairness of a galaxy sample is assessed by statistical functions, and one cannot claim a general fair sample exists without specifying the used statistical method. It is possible that a galaxy sample is fair for one statistical function but not for another function. With our measurements, we conclude that the current SDSS is not able to provide reliable 2PCFs (both for redshift space and projections), PVDs of samples with characteristic luminosity fainter than L^*, or third-order statistics in the weakly nonlinear regime for nearly all volume-limited samples.

For faint volume-limited subsamples, probably due to their very shallow depths, measurements suffer from greater finite volume effects, such that enlarging sky coverage has a larger influence on measurements of statistics than for bright subsamples. The inconsistency observed is a manifestation of cosmic variance due to insufficient sample volume. The variances are comparable to the 1σ jackknife error bars, which are usually regarded as good and robust approximations to the true error bars (Zehavi et al. 2002). Now it seems that the technique underestimates the true variance, with the corresponding results about the habitation of faint galaxies in halos withdrawn from clustering analysis, e.g. Li et al. (2007) and The SDSS Collaboration et al. (2010) were not very concrete. Conclusions about faint galaxies utilizing a galaxy group catalog constructed from SDSS DR4 (Yang et al. 2007, 2008) might also be problematic, so we conjecture that a new group catalog from DR7 may provide a very different paradigm.

In our analysis, PVDs are derived under a general assumption that galaxy pairwise velocities closely follow an exponential distribution. The assumption might not be exact for satellite galaxies, for which the pairwise velocity distribution can be better described by a Gaussian (Tinker 2007). For galaxies with low luminosity, they are most likely satellites; the obtained σ_{12} based on an exponential distribution is biased and so is the relation of PVDs with galaxy luminosity presented in Figure 8. Nevertheless, our PVDs with different versions of VL1 are biased in the same way, and the systematical bias will not affect our basic conclusion that PVD of VL1 from DR7 is very different from that of DR4.

Recently, there have been several works applying 3PCF of SDSS (e.g. Sefusatti et al. 2006; Kulkarni et al. 2007; Marín et al. 2008; Marin 2010; McBride et al. 2010), either to help determining cosmological parameters and galaxy biasing or to diagnose models of galaxy formation. Some results use measurements of volume-limited subsamples of the main SDSS galaxy catalogs in the weakly nonlinear regime. Our analysis however points out that one needs to be very cautious in accepting the relevant conclusions.

Another problem worthy of more discussion is the first zero-crossing point of 2PCF. Of course, part of the problem arises from the finite volume of samples, or at least the problem that the integral constraint is a serious systematic shortcoming for low luminosity galaxy subsamples. However, for subsamples with a large volume of bright galaxies, the absence of anti-correlation at large scales is still puzzling. Instead of criticizing the validity of ΛCDM models, it is probably better to scrutinize the stochastic bias in models of galaxy 2PCF. The halo model alone cannot solve this problem since, at large scales, 2PCF in the halo model boils down to simple multiplication of the bias parameter with

linear 2PCF of dark matter. In the set of parameters used in cosmological application, galaxy 2PCF at large scales by default is fully described by linear bias parameter and 2PCF of dark matter, and the single bias parameter is largely degenerate with some other parameters, such as the normalization of density fluctuation σ_8 and the matter density parameter $\Omega_{\rm m}$. It is unclear if the present estimation of cosmological parameters is significantly biased by the ignorance of possible exotic bias (e.g. the proposal of Coles & Erdogdu 2007).

Acknowledgements This work is funded by the National Science Foundation of China (NSFC) under grant Nos.10643002, 10633040, 10621303, 10873035, and 10725314 and the National Basic Research Program of China (973 program, No. 2007CB815402). The authors have enjoyed meaningful discussions with Xiaohu Yang, Weipeng Lin and Xi Kang.

This publication also makes use of the *Sloan Digital Sky Survey* (SDSS). Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England. The SDSS web site is *http://www.sdss.org/*.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory and the University of Washington.

This publication also made use of NASA's Astrophysics Data System Bibliographic Services.

References

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2006, ApJS, 162, 38
Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, ApJS, 175, 297
Antal, T., Sylos Labini, F., Vasilyev, N. L., & Baryshev, Y. V. 2009, Europhysics Letters, 88, 59001
Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587
Bernstein, G. M. 1994, ApJ, 424, 569
Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, AJ, 129, 2562
Coles, P., & Erdogdu, P. 2007, Journal of Cosmology and Astroparticle Physics, 10, 7
Colless, M., Peterson, B. A., Jackson, C., et al. 2003, ArXiv:astro-ph/0306581
Cooray, A. 2006, MNRAS, 365, 842
Croton, D. J., Springel, V., White, S. D. M., et al. 2006, MNRAS, 365, 11
Efstathiou, G. 1988, in Comets to Cosmology, Lecture Notes in Physics, Berlin Springer Verlag, vol. 297, ed. A. Lawrence, 312
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560
Fry, J. N., & Gaztanaga, E. 1993, ApJ, 413, 447
Gott, J. R., III, Jurić, M., Schlegel, D., et al. 2005, ApJ, 624, 463
Hamilton, A. J. S. 1992, ApJ, 385, L5
Hamilton, A. J. S. 1993, ApJ, 417, 19
Jing, Y. P., Mo, H. J., & Boerner, G. 1998, ApJ, 494, 1
Kaiser, N. 1987, MNRAS, 227, 1

Kang, X., Jing, Y. P., Mo, H. J., & Börner, G. 2005, ApJ, 631, 21
Kazin, E. A., Blanton, M. R., Scoccimarro, R., et al. 2010, ApJ, 710, 1444
Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18
Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, ApJ, 609, 35
Kulkarni, G. V., Nichol, R. C., Sheth, R. K., et al. 2007, MNRAS, 378, 1196
Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
Li, C., Jing, Y. P., Kauffmann, G., et al. 2007, MNRAS, 376, 984
Li, C., Kauffmann, G., Jing, Y. P., et al. 2006, MNRAS, 368, 21
Lupton, R. 1993, Statistics in Theory and Practice
Marin, F. 2010, ArXiv:1011.4530
Marín, F. A., Wechsler, R. H., Frieman, J. A., & Nichol, R. C. 2008, ApJ, 672, 849
Martínez, V. J., Arnalte-Mur, P., Saar, E., et al. 2009, ApJ, 696, L93
McBride, C. K., Connolly, A. J., Gardner, J. P., et al. 2010, ArXiv:1012.3462
Nichol, R. C., Sheth, R. K., Suto, Y., et al. 2006, MNRAS, 368, 1507
Pan, J., & Szapudi, I. 2005, MNRAS, 362, 1363
Pan, J., & Zhang, P. 2010, Journal of Cosmology and Astroparticle Physics 8, 19
Peebles, P. J. E. 1993, Principles of Physical Cosmology
Percival, W. J., Reid, B. A., Eisenstein, D. J., et al. 2010, MNRAS, 401, 2148
Reid, B. A., Percival, W. J., Eisenstein, D. J., et al. 2010, MNRAS, 404, 60
Sefusatti, E., Crocce, M., Pueblas, S., & Scoccimarro, R. 2006, Phys. Rev. D, 74, 023522
Slosar, A., Seljak, U., & Tasitsiomi, A. 2006, MNRAS, 366, 1455
Sylos Labini, F., Vasilyev, N. L., Baryshev, Y. V., & López-Corredoira, M. 2009, A&A, 505, 981
Tegmark, M., Blanton, M. R., Strauss, M. A., et al. 2004, ApJ, 606, 702
The SDSS Collaboration, Zehavi, I., Zheng, Z., et al. 2010, ArXiv:1005.2413
Tinker, J. L. 2007, MNRAS, 374, 477
Yang, X., Mo, H. J., & van den Bosch, F. C. 2003, MNRAS, 339, 1057
Yang, X., Mo, H. J., & van den Bosch, F. C. 2008, ApJ, 676, 248
Yang, X., Mo, H. J., van den Bosch, F. C., et al. 2007, ApJ, 671, 153
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Zehavi, I., Blanton, M. R., Frieman, J. A., et al. 2002, ApJ, 571, 172
Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2005, ApJ, 630, 1
Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791