A NEW METHOD TO CORRECT FOR FIBER COLLISIONS IN GALAXY TWO-POINT STATISTICS

HONG GUO$^1$, IDIT ZEHAVI$^1$, AND ZHENG ZHENG$^2$

$^1$ Department of Astronomy, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA
$^2$ Department of Physics and Astronomy, University of Utah, 115 South 1400 East, Salt Lake City, UT 84112, USA

Received 2011 November 24; accepted 2012 July 11; published 2012 August 21

ABSTRACT

In fiber-fed galaxy redshift surveys, the finite size of the fiber plugs prevents two fibers from being placed too close to one another, limiting the ability to study galaxy clustering on all scales. We present a new method for correcting such fiber collision effects in galaxy clustering statistics based on spectroscopic observations. The target galaxy sample is divided into two distinct populations according to the targeting algorithm of fiber placement, one free of fiber collisions and the other consisting of collided galaxies. The clustering statistics are a combination of the contributions from these two populations. Our method makes use of observations in tile overlap regions to measure the contributions from the collided population, and to therefore recover the full clustering statistics. The method is rooted in solid theoretical ground and is tested extensively on mock galaxy catalogs. We demonstrate that our method can well recover the projected and the full three-dimensional (3D) redshift-space two-point correlation functions (2PCFs) on scales both below and above the fiber collision scale, superior to the commonly used nearest neighbor and angular correction methods. We discuss potential systematic effects in our method. The statistical correction accuracy of our method is only limited by sample variance, which scales down with (the square root of) the volume probed. For a sample similar to the final SDSS-III BOSS galaxy sample, the statistical correction error is expected to be at the level of 1% on scales $\sim 0.1$–$30 h^{-1}$ Mpc for the 2PCFs. The systematic error only occurs on small scales, caused by imperfect correction of collision multiplets, and its magnitude is expected to be smaller than 5%. Our correction method, which can be generalized to other clustering statistics as well, enables more accurate measurements of full 3D galaxy clustering on all scales with galaxy redshift surveys.

Key words: cosmology: observations – cosmology: theory – galaxies: distances and redshifts – galaxies: halos – galaxies: statistics – large-scale structure of Universe

Online-only material: color figures

1. INTRODUCTION

With fiber-fed spectroscopic surveys, such as the Las Campanas Redshift Survey (Shectman et al. 1996), the 2dF Galaxy Redshift Survey (2dFGRS; Colless 1999), the Sloan Digital Sky Survey (SDSS; York et al. 2000; Stoughton et al. 2002; Strauss et al. 2002), and the Galaxy and Mass Assembly Survey (GAMA; Driver et al. 2011; Robotham et al. 2010) can efficiently cover a large sky area and obtain redshifts for a large set of targeted galaxies simultaneously. A well-known problem of using fibers is that the finite size of the fiber plugs prevents two fibers from being placed too close to one another on the same plate. Consequently, a significant fraction of targeted galaxies from a photometric catalog cannot be assigned fibers and obtain measured spectroscopic redshifts. This problem is partly alleviated by having some regions on the sky covered by overlapping plates, but it still results in a fraction of targeted galaxies left with no measured spectroscopic redshifts (e.g., $\sim 7\%$ in the SDSS). These fiber-collided galaxies are a hindrance to any galaxy clustering study. In this paper, we propose and test a new method to account for fiber collision effects and to accurately measure galaxy clustering statistics on small and intermediate scales.

The angular fiber-collision scale, under which two fibers on the same plate collide with each other, is determined by the fiber placement hardware and differs from survey to survey. For SDSS-I and II, this scale is 55′′, corresponding to about 0.1 $h^{-1}$ Mpc at the median redshift $z \sim 0.1$. For SDSS-III, the angular fiber-collision scale is slightly larger, 62′′, corresponding to about 0.4 $h^{-1}$ Mpc (comoving) at the larger median redshift $z \sim 0.55$ of the Baryon Oscillation Spectroscopic Survey (BOSS; Aihara et al. 2011; Eisenstein et al. 2011; White et al. 2011; Anderson et al. 2012). If fiber collisions are not corrected for, galaxy clustering on scales below the collision scale would not be accurately measured. For example, in the case of galaxy two-point correlation functions (2PCF), the effect is seen as a significant decline in the clustering signal below the collision scale (Jing et al. 1998; Yoon et al. 2008). At a fixed angular collision scale, the comoving scale increases with redshift, making the fiber collision a more severe problem in studying small-scale clustering for surveys at higher redshifts. Furthermore, fiber collisions have a non-negligible effect on galaxy clustering measurements even on scales larger than the collision scale (Zehavi et al. 2002, 2005). Therefore, to study galaxy clustering on small scales and to have precise measurements on larger scales, the effect induced by fiber collision has to be corrected for.

In general, there are two approaches to correct the fiber collision effect in measuring galaxy 2PCFs. One is to recover the redshifts of the fiber-collided galaxies and the other is to reconstruct the correct galaxy pair counts. For the former one, a commonly adopted method is to assign each collided galaxy the redshift of its nearest angular neighbor (nearest neighbor correction; e.g., Zehavi et al. 2002, 2005; Berlind et al. 2006). The method has been applied to measure the projected 2PCF $w_p(r_p)$ at a transverse separation $r_p$ for the SDSS-I and II Main galaxy sample (Zehavi et al. 2002, 2005, 2011), which proves to work well down to the fiber collision scale $\sim 0.1 h^{-1}$ Mpc. However, the method fails below the collision scale and does not give a satisfactory correction for the redshift-space correlation...
function. For surveys at higher redshifts, the increase in the (comoving) fiber collision scale limits the application of the nearest neighbor correction to clustering at larger scales.

For the approach of reconstructing the pair counts, a common method is to make use of the angular 2PCF of the spectroscopic sample, \( w_{p}(\theta) \), and that of the parent photometric sample, \( w_{p}(\theta) \), at angular separation \( \theta \), with the same angular and redshift selection applied to both samples (Hawkins et al. 2003; Li et al. 2006a, 2006b; Ross et al. 2008; White et al. 2011). The projected galaxy pair count in computing \( w_{p}(r_{p}) \) is weighted by the ratio \( \tilde{F}(\theta) = [1 + w_{p}(\theta)]/[1 + w_{c}(\theta)] \), where \( \theta \) corresponds to \( r_{p} \) at the median redshift of the survey. Such a weighting scheme incorporates the angular information of the missing galaxies and retrieves the correct angular counts. However, \( w_{p}(r_{p}) \) and \( w_{p}(\theta) \) are not completely correspondent. Li et al. (2006b) test this method with mock catalogs of SDSS galaxies. The method works reasonably well on large scales, but on scales of \( 0.05-1 h^{-1} \) Mpc, the corrected \( w_{p}(r_{p}) \) shows a clear deficit with respect to the true \( w_{p}(r_{p}) \).

Other methods aimed at overcoming the fiber collision problem to measure the small-scale galaxy clustering broadly fall into the above two categories or some combination. For example, the small-scale clustering can be inferred from the full photometric sample by cross-correlating with the spectroscopic sample (Eisenstein et al. 2005a; Masjedi et al. 2006; Watson et al. 2010, 2012; Jiang et al. 2011; Wang et al. 2011). Under the assumption of isotropic clustering, the photometric objects near a spectroscopic target can be considered to be at the same redshift as the target, similar to the nearest neighbor method, and the small-scale clustering can be measured. Contamination from interlopers reduces the signal to noise and needs to be removed statistically. As with the nearest neighbor method, this cross-correlation technique does not work in measuring the full three-dimensional (3D) clustering in redshift space.

Accurate measurements of galaxy clustering are important in many applications. In particular, the small-scale clustering can be used to probe the spatial distribution of galaxies inside the host dark matter halos (e.g., Watson et al. 2010, 2012), to infer the kinematics of galaxies in halos, and to reveal environmental effects on galaxy formation and evolution. In this paper, we propose a new and efficient correction method for the fiber collision effect based on the spectroscopic galaxy sample. The method is proven to work well in the case of measuring both the projected galaxy 2PCFs and the full 3D redshift-space 2PCFs. After describing the mock catalogs used for testing the method and our measurements in Section 2, we present our correction method and explain its theoretical basis in Section 3. Tests of the method with mock catalogs and comparisons with other methods are presented in Section 4. We discuss possible systematic effects of the method in Section 5 and conclude in Section 6.

2. MOCK CATALOGS AND CLUSTERING MEASUREMENTS

Throughout the paper, our proposed new method will be tested and compared to other methods using clustering measurements performed on available realistic mock catalogs. We use the LasDamas galaxy mock catalogs (C. McBride et al. 2012, in preparation), which are constructed by populating galaxies into dark matter halos identified in the LasDamas simulations. The \( N \)-body LasDamas simulations adopt a spatially flat \( \Lambda \)CDM cosmology, with a matter density parameter \( \Omega_{m} = 0.25, baryon density of \( \Omega_{b} = 0.04, \sigma_{8} = 0.8 \) (the primordial matter fluctuation amplitude on scales of \( 8 h^{-1} \) Mpc, linearly extrapolated to \( z = 0 \)), a primordial matter fluctuation spectra index \( n_{s} = 1 \), and a Hubble constant of \( h = 0.7 \). Dark matter halos are identified using a friends-of-friends algorithm (e.g., Davis et al. 1985) with a linking length of \( 0.156 \) times the mean particle separation. Dark matter halos are populated with galaxies through a halo occupation distribution (HOD) approach (e.g., Berlind & Weinberg 2002; Cooray & Sheth 2002; Zheng et al. 2005) and the HOD parameters are determined through modeling the clustering of the early BOSS \( z \sim 0.5 \) CMASS sample (White et al. 2011). Redshift-space distortions are also included in the mock catalogs by accounting for the peculiar velocities of galaxies. The radial and angular selection functions in the mock catalogs are constructed to be uniform.

In total, we make use of 40 LasDamas mock galaxy catalogs. In addition to matching the clustering of CMASS galaxies, the mocks also reproduce the geometry of the early BOSS data. As in White et al. (2011), the LasDamas mock catalogs for CMASS sample are divided into three separate regions. For simplicity, we only use the so-called region B in White et al. (2011), since it has the largest volume and number of galaxies. Each mock of this region consists of about 50,000 galaxies, with a sky coverage of \( \sim 600 \) deg\(^2\) and a redshift range of \( 0.4 < z < 0.6 \), corresponding to a volume of \( \sim 0.16 h^{-3} \) Gpc\(^3\). The mock catalogs do not have the specific tiling mask and fiber collisions imposed on them, and we do that ourselves for our tests as described below.

In this paper, we focus our discussion on the 2PCFs and related statistics. We use the Landy–Szalay estimator (Landy & Szalay 1993) to measure the 2PCFs of galaxies in the mock catalogs,

\[
\xi = \frac{DD - 2DR + RR}{RR},
\]

where \( DD, DR, \) and \( RR \) are the data–data, data–random, and random–random pair counts measured from the data of \( N \) galaxies and random samples consisting of \( N_{R} \) random points. These pair counts are normalized by dividing by \( N(N-1)/2, N N_{R}, \) and \( N_{R}(N_{R}-1)/2, \) respectively.

We measure the 3D \( \xi(r_{p}, \pi) \) and \( \xi(s, \mu) \) functions and the redshift-space 2PCF \( \xi(s) \), where \( r_{p} \) and \( \pi \) are the separations of galaxy pairs perpendicular and parallel to the line of sight (Fisher et al. 1994; Zehavi et al. 2002), \( s = r_{p}^{2} + \pi^{2} \), and \( \mu = \pi/s \) is the cosine of the angle between \( s \) and \( \pi \). The redshift-space 2PCF differs from the real-space one because of the redshift distortion effect induced by galaxy peculiar velocity. The redshift distortion can be mitigated by projecting the 2PCF along the line-of-sight direction, with the projected 2PCF \( w_{p}(r_{p}) \) (Davis & Peebles 1983) defined and measured as

\[
w_{p}(r_{p}) = 2 \int_{0}^{\infty} \xi(r_{p}, \pi)d\pi = 2 \sum_{i} \xi(r_{p}, \pi_{i}) \Delta \pi_{i},
\]

where \( \pi_{i} \) and \( \Delta \pi_{i} \) are the \( i \)th bin of the line-of-sight separation and its corresponding bin size.

Following Hamilton (1992), the redshift-space 2PCF \( \xi(s, \mu) \) can be written in the form of multipole expansion,

\[
\xi(s, \mu) = \sum_{l} \xi_{l}(s) P_{l}(\mu),
\]

where \( P_{l} \) is the \( l \)th order Legendre polynomial. The multipole moment \( \xi_{l} \) is determined by

\[
\xi_{l}(s) = \frac{2l+1}{2} \int_{-1}^{1} \xi(s, \mu) P_{l}(\mu)d\mu.
\]
In linear theory, only the moments of $l = 0, 2, 4$ are non-zero. The monopole $\xi_0(s)$, quadrupole $\xi_2(s)$, and hexadecapole $\xi_4(s)$ are useful for the study of redshift distortions and for obtaining constraints on cosmological parameters (see, e.g., Hamilton 1992; Cole et al. 1994; Tinker et al. 2006; Padmanabhan & White 2008; Kazin et al. 2012; Reid & White 2011). We will also test how well our method recovers these moments.

For measuring the 2PCFs, we use a binning scheme of $\Delta \log r_p = 0.2$, $\Delta \sigma = 1 h^{-1}$ Mpc, $\Delta \log s = 0.2$, and $\Delta \mu = 0.1$. Our results are insensitive to these choices. To obtain the projected 2PCF $w_\ell (r_p)$, $\xi_\ell (r_p, \pi)$ is summed along the line-of-sight direction up to $\pi_{\text{max}} = 40 h^{-1}$ Mpc. Integrating to a larger line-of-sight separation or using realistic angular and redshift selection functions will not affect the conclusion on our correction method. We also note that while the mock catalogs we use were constructed for studying the BOSS CMASS sample, our general conclusions regarding the validity of our method are not dependent on that and would hold for other mock and real data sets.

3. THE NEW CORRECTION METHOD

3.1. Sample Division

In modern galaxy redshift surveys, a tiling algorithm is usually applied to design and place spectroscopic plates (tiles) to cover the survey area. These tiles partially overlap over some of the observed region. In such overlap regions, a galaxy with no fiber assigned in one tile can have a fiber allocated in the other one. For example, in SDSS-I and II, about 40% of the survey area is covered by more than one tile, which eliminates most of the fiber collisions in those regions.

The basic idea of our method of dealing with the fiber collision effect is simply to estimate the contribution of the fiber-collided galaxies to the clustering by using the information in the tile overlap regions. In order to make such an estimate, before measuring any clustering statistic, we divide the full galaxy target sample (i.e., the input photometric sample) into two distinct populations:

Population 1. A subsample in which each galaxy is not angularly collided with any other galaxy in this subsample. We maximize the number of galaxies that are not collided with each other. Such a set of “decollided” galaxies provide a “clean” subsample with no fiber collision correction to be considered.

Population 2. A subsample including all the galaxies that are not in Population 1. This is the set of potentially collided galaxies, and all the fiber-collided galaxies come from this subsample. Each galaxy in this population is within the fiber collision scale of a galaxy in Population 1.

The division of Populations 1 and 2 follows the scheme of assigning fibers in SDSS observations (Blanton et al. 2003). The specific tiling algorithm always recovers one galaxy from collided pairs, and two galaxies from the collided triples if the angular distance of these two galaxies is larger than the fiber collision scale. With our division, Population 1 galaxies always have fibers allocated and have spectroscopic redshifts measured. Some fraction of Population 2 galaxies can also have fibers allocated in the tile overlap regions. Also with our division, in each “collided” close pair of galaxies, one galaxy (with measured redshift) will always be part of Population 1 and the other (with a measured redshift or not) will be part of Population 2. This definition aims to ensure that the pair counts involving the galaxies with redshifts in Population 2 (hereafter, we refer to them as “resolved” galaxies) can be regarded as a representative subset of the overall pair counts. The specific tiling and fiber assignment constraints can make the situation non-trivial if the “representative” assumption is not satisfied in real observation, and we discuss the possible systematics introduced in such a case in Section 5.

We assume that in the survey, we have $N = N_1 + N_2$ galaxy targets, where $N_1$ and $N_2$ are the numbers of galaxies in Populations 1 and 2, respectively. Again, all galaxies in Population 1 have spectroscopic redshifts, and because of fiber collisions only a fraction of Population 2 galaxies do (e.g., in tile overlap regions). We denote the set of targeted galaxies as $D$ (with a total number $N$) and use $D_1$ and $D_2$ to represent the sets of Populations 1 and 2 galaxies in the sample (with numbers $N_1$ and $N_2$, respectively). We use $D'$ and $D_2'$ to denote the corresponding galaxy sets that have spectroscopic redshifts measured (with numbers $N'$ and $N_2'$). Note that by definition, $D_1$ is the same as $D_1$.

The data–data pairs counts can then be decomposed as

$$N_{DD} = N_{D_1D_1} + N_{D_1D_2} + N_{D_2D_2},$$

and

$$N_{D'D'} = N_{D_1D_1'} + N_{D_1D_2'} + N_{D_2D_2'}. \tag{6}$$

The actual numbers of pairs in Equation (5) are what are needed to estimate galaxy 2PCFs, while the pair numbers in Equation (6) are what one obtains in the spectroscopic sample. Our method is to make use of the pair counts of Populations 1 and 2 galaxies to estimate galaxy 2PCFs, while the pair numbers in Equation (6) can also be decomposed in a similar way.

3.2. The Simplified Case

To illustrate our fiber-collision correction method, we first consider the simplest case where Population 2 galaxies are randomly selected to be assigned fibers. In this case, $D_2'$ is a random subset of $D_2$, and the pair counts $N_{D_1D_2}'$ and $N_{D_2D_2}'$ are simply proportional to the full counts $N_{D_1D_2}$ and $N_{D_2D_2}$, i.e.,

$$N_{D_1D_2} = \frac{N_2}{N_2'} N_{D_1D_2'}, \quad N_{D_2D_2} = \left(\frac{N_2}{N_2'}\right)^2 N_{D_2D_2'}. \tag{7}$$

For simplicity, $N_2$ and $N_2'$ are assumed to be large so that $N_2/(N_2 - 1)$ and $N_2'/N_2'$ are replaced with $N_2^2$ and $N_2'^2$, respectively. The above equations provide the way to correct the pair counts obtained from the spectroscopic sample.

Similarly, the data–random pair counts can be corrected as

$$N_{D_1R} = \frac{N_2}{N_2'} N_{D_1R}. \tag{8}$$

Since we can have a large number of random points, the correction here for $N_{D_1R}$ is less noisy than those for $N_{D_1D_2}$ and $N_{D_2D_2}$.

The full pair counts are reconstructed as

$$N_{DD} = N_{D_1D_1} + \frac{N_2}{N_2'} N_{D_1D_2} + \left(\frac{N_2}{N_2'}\right)^2 N_{D_2D_2'} \tag{9}$$

3
from the different populations, Equation (9). The fiber collisions are artificially imposed on catalogs for the simplified case, decomposed according to the mocks with an overall fraction of $D_1$.

The Astrophysical Journal

constraint, determined by the highest redshifts in the mocks.

pair counts. The shaded areas are the 1σ obtained from the spectroscopic galaxies, while the squares show the actual corrected pair counts, obtained from the spectroscopic samples, while the squares show the actual pair counts. The shaded areas are the 1σ error distribution of the corrected $N_{DD}$. The corrected pair counts are further decomposed as the contributions from the different populations, $N_{D_1D_1}$, $N_{D_1D_2}$, and $N_{D_2D_2}$, shown as the dotted lines, labeled here simply as 11, 12, and 22. The vertical dashed lines denote the physical fiber collision scale corresponding to the fiber collision angular constraint, determined by the highest redshifts in the mocks.

$$N_{DD} = N_{D_1D_1} + N_{D_1D_2} + N_{D_2D_2}.$$  (10)

Figure 1 shows the data–data pair counts from the mock catalogs for the simplified case, decomposed according to Equation (9). The fiber collisions are artificially imposed on the mocks with an overall fraction of $N_{D_2}/N_2 = 0.42$ as in the BOSS CMASS catalogs. The solid and dotted curves are the corrected pair counts, obtained from the spectroscopic samples, averaged over the 40 mocks, while the squares show the actual full pair counts. The shaded region in each panel denotes the 1σ scatter from the 40 mock catalogs of the $DD$ pair count (estimated from $DD'$ based on Equation (7)). Comparing the solid curve (plus the small shaded region) with the squares, we see that our correction method accurately recovers the true pair counts over all scales probed, for both $\xi(s)$ and $\xi(r_p, \pi)$. The increase in the scatter on small scales is caused by shot noise, since the numbers of $D_1D_1'$ and $D_2D_2'$ pairs are less than $D_1D_2$ and $D_2D_2'$ pairs.

From Figure 1, we see that there are almost no $D_1D_2$ pairs with separation smaller than $\sim 0.3 \, h^{-1} \, \text{Mpc}$, which is the minimum fiber collision scale (corresponding to the lowest redshift $z = 0.4$). On small scales, the pair count is dominated by $D_1D_2$ pairs, while $D_2D_2$ pairs make a small but non-negligible contribution. Across the fiber collision scale ($\sim 0.3–0.5 \, h^{-1} \, \text{Mpc}$), the dominant contribution to the total pair count shifts from $D_1D_2$ pairs to $D_1D_1$ pairs, reflecting the change from collided to decollided galaxies.

3.3. Theoretical Basis

Our correction method can be put in terms of a decomposition of the 2PCF. With a galaxy sample divided into subsamples, such as red and blue galaxies, central and satellite galaxies, or in our case Populations 1 and 2 galaxies, the 2PCF can be decomposed into contributions from the two-point auto- and cross-correlation functions of subsample galaxies (Zu et al. 2008). This is a fully equivalent way of describing the pair decomposition. In the case of dividing the galaxy sample into Populations 1 and 2 subsamples, we have

$$N_2^2 \xi = N_1^2 \xi_{11} + 2N_1N_2 \xi_{12} + N_2^2 \xi_{22}$$  (11)

$$\approx N_1^2 \xi_{11} + 2N_1N_2 \xi_{12} + N_2^2 \xi_{22},$$  (12)

where $\xi$ can be either $\xi(s)$, $\xi(r_p, \pi)$, or $w_p(r_p)$, and $\xi_{11}$, $\xi_{12}$, and $\xi_{22}$ are the two-point auto-correlation function of Population 1 galaxies, cross-correlation function between Population 1 and Population 2 galaxies, and auto-correlation function of Population 2 galaxies, respectively. The second line in the equation uses the component correlation functions $\xi_{11}$ and $\xi_{22}$ estimated from the $D_1$ and $D_2$ galaxies in the spectroscopic sample to approximate $\xi_{12}$ and $\xi_{22}$, which is the key point of our correction method.

If $D_1$ galaxies are representative of $D_2$ galaxies, i.e., $D_1^\prime$ is a random subset of $D_2$, the ensemble average of $\xi_{12}$ ($\xi_{22}$) should equal to that of $\xi_{11}$ ($\xi_{22}$).

$$\langle \xi_{12} \rangle = \langle \xi_{11} \rangle, \quad \langle \xi_{22} \rangle = \langle \xi_{22} \rangle.$$  (13)

That is, with ensemble averages, the approximation symbol can be replaced with the equals sign. Therefore, our method using Equation (12) provides an unbiased correction, which is a desired merit. In practice, the method is only applied to one realization in the ensemble, so differences are expected between the true underlying 2PCF and the 2PCF corrected with our method. However, any discrepancies would only result from the fact that a smaller number of galaxies are used to estimate the component 2PCFs. They show up as sample variance, not systematic errors. Potential systematics caused by the violation of the “representative” assumption of $D_1^\prime$ galaxies are discussed in Section 5.

The component 2PCFs can be estimated with the Landy–Szalay estimator,

$$\xi_{11} = \frac{D_1D_1 - 2D_1R_1 + R_1R_1}{R_1R_1},$$  (14)

$$\xi_{12} = \frac{D_1D_2 - D_1R_2 - D_2R_1 + R_1R_2}{R_1R_2},$$  (15)

$$\xi_{22} = \frac{D_2D_2 - 2D_2R_2 + R_2R_2}{R_2R_2}.$$  (16)

In the case that $D_1^\prime$ is a random subset of $D_2$, we can use the same random sample, i.e., $R_1 = R_2 = R$ in the above equations. It is not surprising that when substituting the above equations into Equation (12), we end up with exactly the same result as with Equations (9) and (10).

3.4. The Tiled Case

Taking into account the tiling in the real observational situation would change the geometry of the distribution of $D_1^\prime$ galaxies such that most of them would preferentially populate the tile overlap regions. The survey geometry can be described
by the individual sectors, defined by areas of sky covered by unique sets of tiles (Blanton et al. 2003). We use \( N_{\text{tile}} \) to denote the number of tiles covering each sector. For \( N_{\text{tile}} = 1 \) regions, essentially no fiber-collided galaxies (\( D_2 \)) can have spectroscopic redshifts. For \( N_{\text{tile}} \geq 2 \) regions, however, most of these will be resolved, with spectroscopic redshifts measured as a result of repeated observations in these tile overlap regions.

In practice, some fiber-collided \( D_2 \) galaxies in \( N_{\text{tile}} = 1 \) regions may still have spectroscopic redshifts from other surveys. In the case of the CMASS sample, about 5.5% of all the galaxies do not have redshifts. Specifically, the values of the ratio \( N_{\text{tile}}^2/N_2 \) are about 5%, 71%, and 87% in \( N_{\text{tile}} = 1, 2, \) and 3 regions, respectively. The 5% recovered galaxies in \( N_{\text{tile}} = 1 \) regions are included from the “Legacy survey” of SDSS-I and II (Eisenstein et al. 2011).

In order to mimic the full observational case, we impose the tile placement of the BOSS survey on all mock catalogs as well as fiber collisions according to the tiling algorithm with the appropriate \( N_{\text{tile}}^2/N_2 \) ratios specified above. We stress that the validity of our method does not depend on the specific values of these parameters, and these are simply chosen here to resemble the values in the BOSS survey. Figure 2 illustrates the galaxy distribution and the tile placement in a small section of one of our mocks. The open symbols are all \( D_1 \) galaxies. Filled squares denote \( D_2 \) galaxies, where blue ones mark resolved galaxies (\( D_2^{\bullet} \)) and red ones are those galaxies in \( D_2 \) whose redshifts are missing due to fiber collisions. The \( D_2^{\bullet} \) galaxies are not randomly distributed over the whole observed sky—they mostly occupy regions with \( N_{\text{tile}} \geq 2 \).

For the tiled case, our correction method as in Section 3.3 remains the same, and the only modification is the need to account for the specific geometry of the \( D_2^{\bullet} \) galaxy distribution. One straightforward way to do this is to create separate random catalogs \( R_1 \) and \( R_2 \) for \( D_1 \) and \( D_2^{\bullet} \) galaxies. The \( R_1 \) catalog can be created as usual by incorporating the standard radial and angular selection functions of the sample. We note that the latter is commonly characterized and applied as a function of the individual sectors (e.g., Zehavi et al. 2002). For the \( R_2 \) catalog, an additional angular completeness mask needs to be applied, with the number of random points in each individual sector modified by the corresponding \( N_{\text{tile}}^2/N_2 \) ratio. For example, in Figure 2 the circular tile centered at R.A. \( \sim 158^\circ \) and decl. \( \sim -2^\circ \) is composed of eight individual sectors, as a result of tile overlap, each with its own \( N_{\text{tile}}^2/N_2 \) value. In the extreme case that there are no \( D_2^{\bullet} \) galaxies in \( N_{\text{tile}} = 1 \) regions, these regions would be empty in the \( R_2 \) catalog.

Alternatively, one could choose to up-weight \( D_2^{\bullet} \) galaxies in each sector by \( N_{\text{tile}}^2/N_2 \) instead of creating a separate random sample for the \( D_2^{\bullet} \) galaxies. However, for \( N_{\text{tile}} = 1 \) regions, \( N_{\text{tile}}^2/N_2 \) is usually a tiny number, which will introduce large errors when adopting the up-weighting method. Therefore, we prefer to down-weight the random catalog to accurately account for the angular distribution and selection of \( D_2^{\bullet} \) galaxies.

4. TESTING THE METHOD

We test our correction method and compare with other commonly used methods by measuring the 2PCFs and related statistics in the LasDamas mocks. We first show results for the simplified case and then test our method on our tiled mocks.

4.1. The Simplified Case

To apply our correction method, we first divide the mock galaxies into the two populations as discussed in Section 3.1. We first define all galaxies in the sample as \( D_1 \) and proceed as below: (1) for each collided pair in \( D_1 \), we randomly change one galaxy to be \( D_2^{\bullet} \); (2) for each collided triplet (and higher multiplet) in \( D_1 \), we assign as \( D_2^{\bullet} \) the galaxy that collides with the most \( D_1 \) galaxies; and (3) we repeat steps (1) and (2) until none of the galaxies in \( D_1 \) collide with one another. This method maximizes the number of galaxies in \( D_2^{\bullet} \) and mimics the real tiling algorithm. In reality, some objects (e.g., quasars) have higher fiber allocation priority and some assigned fibers do not result in reliable redshifts due to the hardware limitation. This complicates the tiling algorithm, but the targeted galaxies can still be divided into the two populations properly. The population division results in about 9.5% of all galaxies being in the \( D_2^{\bullet} \) sample. We then randomly assign fibers to 42% of the \( D_2^{\bullet} \) galaxies (i.e., \( N_{\text{tile}}^2/N_2 \approx 0.42 \)), so that about 5.5% of all galaxies do not have fibers assigned, mimicking the fiber collision fraction in the CMASS sample. With the simplest case considered here, we only need to create one random catalog.

Figure 3 shows the 2PCFs \( w_2(P) \) and \( \xi(s) \) from our correction method and the comparisons with those from other methods. Filled squares represent the actual 2PCFs, measured using all the mock galaxies. Error bars reflect the 1\( \sigma \) variation among the 40 LasDamas mocks. Open squares are obtained using only those galaxies that have fibers assigned, corresponding to the case without any fiber-collision correction. In each panel, the vertical dotted line is the fiber-collision scale corresponding to the highest redshifts included in the mocks. Not accounting for fiber collisions, the measured 2PCF drops sharply below the collision scale, and it is also underestimated on larger scales. Note that the error bars in the top panels represent the variance (fluctuation) among the 40 mocks, and they do not reflect directly the accuracy of our correction method. In the bottom left panel, we compute the ratio \( w_2(P)/w_2(P,\text{true}) \) (and similarly for \( \xi(s) \))
Similar to Figure 1, the contribution from fiber collision recovers well all the components of the correlation corresponding to the fiber collision angular constraint, determined by the highest redshifts in the mocks.

In each panel, the blue solid curve is the 2PCF obtained with our correction method, averaged over the 40 mocks. Error bars are the 1σ scatter among the 40 mocks. Our correction method works very well for both \( w_p \) and \( \xi(s) \) over all measured scales. In particular, it works down to the smallest scales for which we have enough pair counts to estimate the 2PCF \( (r_p \sim 0.1 h^{-1} \text{Mpc}) \). The fractional errors for \( w_p(r_p) \) on small scales are smaller than those for \( \xi(s) \), since for the same value and bin size of \( r_p \) and \( s \) there are more pairs in computing \( w_p(r_p) \) than \( \xi(s) \) due to the projection over large line-of-sight separation. On scales above \( 1 h^{-1} \text{Mpc} \), the errors in both \( w_p \) and \( \xi(s) \) are similar, at the level of \(<3\%\).

The full collision-free 2PCFs and the ones corresponding to our new correction method are further decomposed, in the top panels, into the contributions from the component 2PCFs (dashed and dotted lines; see Equations (11) and (12)). Our correction recovers well all the components of the correlation function. Similar to Figure 1, the contribution from \( \xi_{11} \) dominates on large scales, and it quickly decreases below the fiber collision scale. The contribution from component \( \xi_{12} \) dominates on scales smaller than the collision scale, where that of \( \xi_{22} \) makes a non-negligible contribution. Without any fiber collision correction, the measured 2PCFs (open squares) are still a combination of the three component 2PCFs, but with different coefficients than those in Equation (12). The significant decline at the fiber collision scale in the uncorrected \( w_p(r_p) \) reflects the transition from \( \xi_{11} \) to \( \xi_{12} \) dominated regime. The low amplitude below the collision scale suggests the lack of contribution from \( \xi_{12} \).

For comparison, the nearest neighbor method (red curves) yields the correct \( w_p(r_p) \) above the fiber collision scale, due to the line-of-sight projection, but it clearly overestimates \( w_p(r_p) \) below this scale (see also Zehavi et al. 2002, 2005), because of the increasing importance of the line-of-sight separation on small scales. For \( \xi(s) \), the nearest neighbor estimate fails to recover the correct values over most scales, deviating by more than 1σ for all scales \(<10 h^{-1} \text{Mpc} \). Thus, its use is very limited for redshift-space clustering measurements. The 2PCFs (both \( w_p \) and \( \xi(s) \)) from the angular correction method systematically deviate from the true values except for small scales, and on large scales they approach the estimate with no fiber-collision correction applied (see White et al. 2011). The deviation is at a level \(<10\%\), consistent with the finding of Li et al. (2006b). Physical explanations for the deviations seen in the nearest neighbor and angular corrections are provided in Section 6.

In contrast, our correction method appears unbiased, not showing any systematic errors. The measurement errors are larger than those with all galaxies (top panels in Figure 3). This is easy to understand—our method only uses galaxies with fibers assigned (i.e., with spectroscopic redshifts), which is less than the total number of galaxies and therefore the sample variance increases. A survey of larger volume would help to reduce the sample variance. The other two commonly used correction methods do introduce scale-dependent systematic errors, although the nearest neighbor method does work remarkably well for \( w_p \) measurements on large scales.
4.2. The Tiled Case

To test our method for the more realistic situation, we impose the BOSS tiling geometry on the mocks (see an illustration in Figure 2). The division of $D_1$ and $D_2$ galaxies is still the same as in Section 4.1. By definition, all $D_1$ galaxies have fibers assigned. We then randomly assign fibers to $N'_2/N_2 = 5\%$, 71\%, and 87\% of $D_2$ galaxies in $N_{\text{tile}} = 1$, 2, and 3 regions, respectively. These fractions are consistent with those in the CMASS sample. As with the simplest case, there are about 5.5\% of all galaxies left without fibers.

As explained in Section 3.4, we create separate random catalogs $R_1$ and $R_2$ for $D_1$ and $D'_2$ galaxies, respectively. The $R_1$ catalog has the overall survey geometry of the mock samples. For the $R_2$ catalog, in addition to the overall geometry, we impose a random subsampling of an $N'_2/N_2$ fraction of points in each individual sector. Note that when applying the method on actual data, it is best to use the individual $N'_2/N_2$ values in each sector for constructing $R_2$, rather than the average value for each $N_{\text{tile}}$, to fully account for the distribution of $D'_2$ galaxies.

The 2PCFs’ results from the different correction methods on the tiled mocks are shown in Figure 4. They are quite similar to those seen in Figure 3 for the simplified case. Our correction method still accurately recovers both $w_p(r_p)$ and $\xi(s)$ over all measured scales. The errors increase only slightly, compared to...
the right panels of Figure 4 is the monopole of the 3D redshift-space distortions in the nonlinear regime (e.g., Tinker et al. 2006; Tinker 2007; Reid & White 2011), which has been hindered so far by the fiber collision effect. In contrast, the density effect is not significant in our mocks, but there may exist another density-caused subtle effect in observational data. If the D2 galaxies in the overlap regions and the fiber-collided D2 galaxies are not exactly the same type of galaxies given the environment difference, they might have a systematic difference in clustering. We use the CMASS DR9 sample to test this and find that the D2 and D1 galaxies do have the same color and apparent magnitude distribution, not unexpected given that the tiling algorithm itself is just an angular selection and does not involve any physical properties of the targeting galaxies.

5. SYSTEMATIC EFFECTS

One key assumption in our method presented above is that the D2 galaxies are a representative (or random) subset of D1 galaxies, which ensures that pair counts involving D2 galaxies can be recovered from those involving D1 galaxies in a simple way. In reality, however, the “representative” assumption may not be fully satisfied, given the constraints from tiling and fiber assignment algorithm. There are two types of potential systematic effects in our correction method, when applied to the real observational data. The first is related to the possible difference in the target density between the overlap and non-overlap regions, due to specifics of the tiling algorithm. The second is related to galaxy pairs in collided triplets (or higher-order collided groups). In what follows, we discuss these two effects and provide solutions.

5.1. Density Effect

If the tiling algorithm is optimized to assign the most fibers to the galaxy targets, it may preferentially place overlapping tiles in higher number density regions (Blanton et al. 2003; J. L. Tinker et al. 2011, private communication). In such a case, D2 galaxies come from regions of slightly higher number density, not necessarily representative of the overall D2 galaxies. This may limit the accuracy of our use of (ξ12) = (ξ12) and (ξ12) = (ξ12). Observationally, the impact of such potential density variations on our method can be evaluated, if we have a complete representative observed area, where all fiber-collided galaxies have been resolved by repeat observations. Theoretically, a set of mocks with the actual tiling algorithm used in observations applied directly to it, can help understand the impact. Since neither the required observed area nor the realistic tiled mocks are yet available, we perform various tests with our mocks to estimate the impact of the density effect.

We define a galaxy density measure of the overlap regions (“overlap density”) as δ ≡ noverlap/nall − 1, where noverlap and nall are the number densities of galaxies in the overlap regions and the whole survey regions, respectively. Then the question becomes whether the accuracy of our recovered 2PCFs depends on the value of δ. While we do not expect systematic density variations in the overlap regions in the ensemble of our tiled mocks, the variation caused by sample variance in individual mocks can be used to study the density effect.

We have calculated the “overlap density” for each of our 40 mock catalogs using the number densities of both all the galaxies and just the D1 ones, which we term as δall and δD1, respectively. Figure 6 shows the dependence of wp/wp,ξ on the two overlap density measures, at different scales rD. In each panel, the open circles are the measurements from the 40 LasDamas mocks and the solid line is the linear least-square fit to the data points. On large scales (rD ≳ 1 h⁻¹ Mpc), we find no dependence of the 2PCF on the overlap density. On small scales (rD ≲ 1 h⁻¹ Mpc), the scatter is large, but the 2PCF ratio is again consistent with no dependence on the overlap density. Overall, no strong systematic dependence on either δall or δD1 is found over the overlap density range probed by our mocks (−0.06 to 0.06). To put this into the context of real observation, we calculate the overlap densities in the CMASS DR9 sample and find them to be δall = 0.032 and δD1 = 0.085. Based on our tests, we expect that the density effect in such a sample is not a concern. We have also verified that the angular 2PCFs of the overlap and non-overlap regions show consistent clustering amplitudes in the CMASS DR9 sample, confirming that the density effect can be likely ignored.

The density effect is not significant in our mocks, but there may exist another density-caused subtle effect in observational data. If D2 galaxies in the overlap regions and the fiber-collided D2 galaxies are not exactly the same type of galaxies given the environment difference, they might have a systematic difference in clustering. We use the CMASS DR9 sample to test this and find that the D2 and D1 galaxies do have the same color and apparent magnitude distribution, not unexpected given that the tiling algorithm itself is just an angular selection and does not involve any physical properties of the targeting galaxies.

As a whole, our investigation suggests that the density difference between the tile overlap regions and the whole survey region does not introduce a noticeable systematic effect in our method.

5.2. Effect of Collision Groups

There is another effect that can violate the “representative” assumption. Galaxies in the D2 population can be associated with different collision groups. For D2 galaxies that are part of collision pairs only (N.B.: all such pairs are D1–D2 or 1–2 pairs given that the number of D1 galaxies is maximized in our division), D2 galaxies in overlap regions are certainly a random subset of such D2 galaxies. There are also D2 galaxies that are...
Dependence of the ratio $w_p/w_p_{\text{true}}$ on the overdensity of all galaxies $\delta_{\text{all}}$ and that of $D_2$ galaxies $\delta_{D_2}$ in tile overlap regions at various scales $r_p$. In each panel, the open circles are the measurements from the 40 LasDamas mocks and the solid line is the linear least-square fit to the data.

(A color version of this figure is available in the online journal.)

In order to assess this, we modify the way of assigning fibers to the galaxies in our mocks to mimic the real tiling algorithm. In Section 4.2, the recovered $D_2$ galaxies in each sector are randomly selected according to the prescribed fraction. We now assign fibers to $D_2$ galaxies in each colliding group according to the number of tiles covering that sector. For example, for three galaxies colliding together in $N_{\text{tile}} = 2$ regions, at most two of them would be assigned fibers, one from $D_1$ and the other from $D_2$.

Figure 7 shows the 2PCFs and the decomposition components for our method without any correction for such a “triplet effect” (green lines) compared with the true ones (open symbols and black dotted lines). It clearly shows that, in this case, due to the missing $D_2$-$D_2$ pairs the $\xi_{22}$ components on scales less than the fiber collision scale are significantly decreased. The $\xi_{22}$ term makes around a 15% contribution to $w_p$ on small scales, so without any correction for this the resulting $w_p$ is now significantly underestimated below the fiber collision scale, and $\xi(s)$ is also affected.

A natural solution to correct for this systematic effect is to extend Equation (11) to three populations—$D_1$, $D_2$, and $D_3$, where $D_2$ here refers to the population of galaxies that collide with $D_1$ but do not collide with each other, and $D_3$ is the rest of the galaxies, corresponding to higher-rank collision groups.
Equations (11) and (12) are then revised as
\[
N^2\xi = N^2_{D_1D_2} + N^2_{D_1D_3} + 2N_1N_2\xi_{12} + 2N_1N_3\xi_{13} + 2N_2N_3\xi_{23}
\]
\[
\approx N^2_{D_1D_2} + N^2_{D_1D_3} + N^2_{D_2D_3} + 2N_1N_2\xi_{12} + 2N_1N_3\xi_{13} + 2N_2N_3\xi_{23}.
\]

We refer to this correction method as “1–2–3 fix” below. The limitation of this correction is the small number of resolved \(D_3\) galaxies, since one needs regions covered by three or more tiles to recover \(D_3\) galaxies, which only occupy a small fraction of the full survey region. Therefore, such a correction method would have large sample variance.

Another way of correcting for the collision triplets is to still use Equation (12) but simply add to \(D_1D_2\) the estimated missing pair counts in the tile overlap regions. The missing 2–2 pairs in 1–2–2 triplets are statistically equivalent to 1–2 pairs in such triplets. We therefore account for the unrepresented \(D_1D_2\) close pairs using the recovered \(D_1D_2\) pairs in these triplets. We denote the \(D_2\) galaxies (the definition of \(D_2\) population here still follows Equation (11)) in 1–2–2 collision triplets as \(D_2\) and the recovered \(D_2\) galaxies in such triplets as \(D_2'\). The total number of \(D_2D_2\) close pairs in each \((r_p, \pi)\) bin below the fiber collision scale in overlap regions can then be written as follows:
\[
N_{D_2D_2} = \left(\frac{N_2^{\text{ol}}}{N_2^{\text{ol}}}\right)^2 N_{D_2D_2'} + \frac{1}{2} \frac{N_1}{N_1'} N_{D_1D_2'}
\]
\[
= \left(\frac{N_2^{\text{ol}}}{N_2^{\text{ol}}}\right)^2 \left[ N_{D_2D_2'} + \frac{1}{2} \frac{N_1}{N_1'} N_{D_1D_2'} \left(\frac{N_2^{\text{ol}}}{N_2^{\text{ol}}}\right)^2 \right],
\]
where \(N_1\) and \(N_1'\) are the numbers of \(D_1\) and \(D_1'\) galaxies in the tile overlap regions, and \(N_2^{\text{ol}}\) and \(N_2^{\text{ol}}\) are the total and recovered number of \(D_2\) galaxies in overlap regions, respectively. The first term in Equation (19) represents the resolved \(D_2D_2\) pairs in the sample (typically only in \(N_{\text{tile}} > 2\) regions), while the second term is an estimate of the missing close pairs (mostly in \(N_{\text{tile}} = 2\) regions). The expression \(N_1/N_1' = 2\), so this simply adds a missing \(D_2D_2'\) pair for each \(D_1D_1'\) pair. The correction term itself is then the second term in the square brackets in Equation (20), adding the normalized missing pairs in 1–2–2 triplets below the fiber collision scale, and we refer to it as the “simple fix.”

Figure 7 also shows the results of applying these two methods to the mocks with the modified fiber assignment. Both corrections alleviate most of the systematics. The “1–2–3” fix is in better agreement with the underlying measurements (in particular for \(\xi(s)\)), but as discussed above, it has large sample variance, leading to a \(\sim50\%\) increase in the error bars. The simple fix reduces the bulk of the systematics, but a few percent deficit remains on small scales. This could be caused by the complicated structure of the high-order colliding groups, remaining systematics associated with the \(D_1D_1'\) pair counts and sample variance.

Although the two correction methods proposed here are not as accurate as in the simple tiled case, they still provide workable estimates of the true 2PCFs. We regard the simple fix as the more practical one, due to its simple application and the increased sample variance of the “1–2–3” fix. We thus advocate simply incorporating the additional term (Equation (20)) into the overall method. We have explored many alternative corrections to the issue of collision groups, with varying complexity and success, and will study it further when a more realistic set of tiled mocks or a complete fiber-collision-free subsample in the

**Figure 7.** Effect of collision triplets on \(w(r_p)\) and \(\xi(s)\), and comparison between the two correction methods. The unresolved \(D_2\) galaxies in collision triplets, if not corrected, lead to an underestimate of the 2PCF on small scales (green). The “1–2–3” fix (red) and the simple fix (blue) methods can help to reduce the effect, and the latter has smaller sample variance. In the top panels, individual components of the 2PCF decomposition are plotted for the true case (black dotted lines) and for our corrections (blue and green dashed lines).

(A color version of this figure is available in the online journal.)
BOSS ancillary program becomes available. We note that, as mentioned above, the LasDamas mocks have in fact $\sim 60\%$ more close triplets than the real CMASS sample. Thus, the magnitude of the effect presented here and any residual systematics are conservative and are likely smaller in the real data.

We also test the influence of the triplet effect in the mocks on the nearest neighbor and angular correction methods, as shown in Figure 8. The nearest neighbor method is mostly unaffected by this, still recovering the correct correlation function on large scales. The angular correction is impacted by the triplet effect, especially for the redshift-space correlation function, though to a lesser extent than our method. Based on our tests, the overall accuracy of our method appears to still be better than the angular correction and nearest neighbor methods.

6. SUMMARY AND DISCUSSION

In this paper, we present a novel method for correcting the effects of fiber collisions in galaxy clustering statistics, utilizing resolved fiber collisions in tile overlap regions. The key element is dividing the target galaxy sample into two distinct populations influenced differently by the fiber collisions and combining their contributions according to Equation (12). The collided galaxies, making up our so-called Population 2 subsample, are partially resolved, mainly in the tile overlap regions. These resolved collided galaxies allow us to recover and measure the clustering statistics on all scales. The distinct spatial distribution of the galaxies in Population 2 which were assigned fibers can be properly accounted for by their angular completeness $N'_{ij}/N_{ij}$ in each individual tile sector. We explain the theoretical basis for our method and extensively test it on realistic mock catalogs. We demonstrate that both the projected and the 3D redshift-space 2PCFs can all be very well recovered, assuming that the recovered Population 2 galaxies are representative. With that assumption, the correction method is accurate and unbiased and limited only by sample variance.

In the real observation, the density variations between overlap and non-overlap regions and the non-random selection of the Population 2 galaxies in higher colliding groups are the main potential systematic effects. While the density effect can be largely ignored as we demonstrate in the mocks, the influence of higher colliding groups should be carefully taken into account in order to provide a reasonable estimate of the true 2PCFs. We proposed a simple fix to the missing close pairs caused by such colliding groups. The remaining systematic effect on small-scale 2PCFs from such a method is expected to be lower than $\sim 5\%$. The exact values of the systematic errors are hard to obtain, given that we are limited by the sample variance on small scales. By shifting the plate placement of the tiling geometry, we find that the deviation from the true 2PCF based on our correction method is likely to be below 3\%. To be conservative, we quote 5\% as an upper limit of the systematic errors for our method.

We also contrast our method to the commonly used nearest neighbor and angular corrections. While these approximations work well for specific statistics on some scales, they are generally not accurate enough to give an unbiased estimate over all scales. For the nearest neighbor correction, assigning the collided galaxy the redshift of its neighbor mainly influences the line-of-sight separation $\pi$, while $r_p$ is only changed because of the non-plane-parallel effect. For projected statistics like $w_p(r_p)$, on scales larger than the fiber-collision scale, the pair counts are dominated by non-collided galaxies, and the nearest neighbor correction works extremely well. On small scales, however, the pair counts are dominated by collided galaxies. The nearest neighbor redshift assignment leads to an overestimate of the number of pairs within the projected separation $\pi_{\text{max}}$ and thus an overestimate of $w_p(r_p)$, which causes the nearest neighbor correction to fail. For $\xi(s)$, the nearest neighbor correction fails below and above the fiber-collision scale, since having the correct line-of-sight separation is more crucial, and the method does worse on all scales smaller than $15 h^{-1} \text{Mpc}$. The nearest
our mocks, so our statistical correction error will be 2.4 times smaller when applied to this sample. The final survey will cover about 10,000 deg², and the correction error in \( w_p \) from sample variance is expected to be \( \sim 1.5\% \) at \( r_p \sim 0.1 \, h^{-1} \text{Mpc} \) and \( 0.6\% \) at \( r_p \sim 30 \, h^{-1} \text{Mpc} \). The residual systematics may also scale down somewhat with the decreased sample variance, and we plan to develop a better treatment of them in future work.

The method thus enables more accurate measurements of galaxy clustering statistics on small and intermediate scales. In particular, it will enable us to reliably extend the measurements to smaller scales than obtained before and to recover the full 3D redshift-space correlation functions on small scales. This will allow a better probe of the distribution of galaxies within halos. HOD modeling of these new measurements will provide new constraints on the spatial and velocity distributions of galaxies inside halos. Measurement and modeling of redshift-space distortions in both the linear and nonlinear regimes will also improve constraints on cosmological parameters. These applications will be explored in future work. We note, however, that as our method is associated with increased sample variance, due to the limited sky coverage of \( D_2 \) galaxies, it might not be ideal for very large scale clustering measurements, such as those performed for measuring the baryon acoustic oscillation signature (Eisenstein et al. 2005b). In such cases, other methods such as the nearest neighbor correction can be considered.

The new method will allow for reliable galaxy clustering measurements in current and future surveys. While we were motivated by upcoming measurements in the SDSS-III BOSS survey, and tested the method on corresponding mock catalogs, the method can be applied to any fiber-fed large surveys such as the SDSS-I and II, 2dFGRS, and planned BigBOSS (Schlegel et al. 2011). We have focused here on two-point auto-correlation functions, but our methodology is more broadly applicable to other related statistics. It can be easily extended to cross-correlation functions (see the Appendix), which would not suffer from the systematic effect of higher-order colliding groups if the two samples of targets come from different surveys. The correction method can be further generalized to higher-order statistics, e.g., the three-point correlation function (Jing & Börner 2004; Kayo et al. 2004; McBride et al. 2011), with the effect of collision groups treated more accurately. The method can certainly be helpful in accurately measuring statistics in Fourier space as well, such as the power spectrum and bispectrum of galaxies.

We thank the anonymous referee for insightful suggestions that significantly improved the paper. We thank Cameron McBride for providing the LasDamas mocks and Chris Mihos for providing access to the High Performance Research Computing Cluster at Case Western Reserve University. We also thank Will Percival, Ariel Sanchez, Ashley Ross, Martin White, Andreas Berlind, Michael Blanton, Daniel Eisenstein, Shirley Ho, Yipeng Jing, Cheng Li, Nikhil Padmanabhan, John Parejko, Jeremy Tinker, David Wake, and David Weinberg for stimulating discussions and the BOSS galaxy clustering working group as a whole for support and encouragement. We thank the LasDamas project for making their mock catalogs available. This work is supported by NSF grant AST-0907947.

APPENDIX

Our new method for fiber-collision correction is also valuable for the measurement of cross-correlation functions. For

...
cross-correlation functions, the key equation (12) is revised accordingly. When cross-correlating two samples $a$ and $b$, both are divided into Populations 1 and 2. The decomposition equation becomes

$$
N_a N_b \xi = N_{a1} N_{b1} \xi_{a1,b1} + N_{a1} N_{b2} \xi_{a1,b2} + N_{a2} N_{b1} \xi_{a2,b1} + N_{a2} N_{b2} \xi_{a2,b2}
$$

(A1)

$$
\approx N_{a1} N_{b1} \xi_{a1,b1} + N_{a1} N_{b2} \xi_{a1,b2} + N_{a2} N_{b1} \xi_{a2,b1} + N_{a2} N_{b2} \xi_{a2,b2}
$$

(A2)

where $N_{a1}$ ($N_{b1}$) and $N_{a2}$ ($N_{b2}$) are the numbers of galaxies in Populations 1 and 2 in sample $a$ ($b$), and $a_1'$ and $b_2'$ denote the subsets of $a_2$ and $b_2$ for which redshifts were obtained, analogous to $D_{a1}'$ and $D_{b2}$.

REFERENCES

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29
Anderson, L., Aubourg, E., Bailey, S., et al. 2012, arXiv:1203.6594
Berlind, A. A., Frieman, J., Weinberg, D. H., et al. 2006, ApJS, 167, 1
Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587
Blanton, M. R., Lin, H., Lupton, R. H., et al. 2003, AJ, 125, 2276
Cole, S., Fisher, K. B., & Weinberg, D. H. 1994, MNRAS, 267, 785
Colless, M. 1999, Phil. Trans. R. Soc., 357, 105
Cooray, A., & Sheth, R. 2002, Phys. Rep., 372, 1
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465
Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971
Eisenstein, D. J., Blanton, M., Zehavi, I., et al. 2005a, ApJ, 619, 178
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005b, ApJ, 633, 560
Fisher, K. B., Davis, M., Strauss, M. A., Yahil, A., & Huchra, J. P. 1994, MNRAS, 267, 927
Hamilton, A. J. S. 1992, ApJ, 385, L5
Hawkins, E., Maddox, S., Cole, S., et al. 2003, MNRAS, 346, 78
Jiang, T., Hogg, D. W., & Blanton, M. R. 2011, ApJ, submitted (arXiv:1104.5483)
Jing, Y. P., & Börner, G. 2004, ApJ, 607, 140
Jing, Y. P., Mo, H. J., & Börner, G. 1998, ApJ, 494, 1
Kazin, E. A., Sánchez, A. G., & Blanton, M. R. 2012, MNRAS, 419, 3223
Li, C., Kauffmann, G., Jing, Y. P., et al. 2006a, MNRAS, 368, 21
Li, C., Kauffmann, G., Wang, L., et al. 2006b, MNRAS, 373, 457
Masjedi, M., Hogg, D. W., Cool, J. R., et al. 2006, ApJ, 644, 54
McBride, C. K., Connolly, A. J., Gardner, J. P., et al. 2011, ApJ, 726, 13
Padmanabhan, N., & White, M. 2008, Phys. Rev. D, 77, 123540
Percival, W. J., Baugh, C. M., Bland-Hawthorn, J., et al. 2001, MNRAS, 327, 1297
Reid, B. A., & White, M. 2011, MNRAS, 417, 1913
Robotham, A. S. G., Driver, S. P., Norberg, P., et al. 2010, PASA, 27, 76
Ross, N., Shanks, T., Cannon, R. D., et al. 2008, MNRAS, 387, 1323
Schlegel, D., Abdalla, F., Abraham, T., et al. 2011, arXiv:1106.1706
Shectman, S. A., Landy, S. D., Oemler, A., et al. 1996, ApJ, 470, 172
Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485
Strass, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810
Tinker, J. L. 2007, MNRAS, 374, 477
Tinker, J. L., Weinberg, D. H., & Zheng, Z. 2006, MNRAS, 368, 85
Wang, W., Jing, Y. P., Li, C., Okumura, T., & Han, J. 2011, ApJ, 734, 88
Watson, D. F., Berlind, A. A., McBride, C. K., Hogg, D. W., & Jiang, T. 2012, ApJ, 749, 83
Watson, D. F., Berlind, A. A., McBride, C. K., & Masjedi, M. 2010, ApJ, 709, 115
White, M., Blanton, M., Bolton, A., et al. 2011, ApJ, 728, 126
Yoon, J. H., Schawinski, K., Sheen, Y.-K., Ree, C. H., & Yi, S. K. 2008, ApJS, 176, 414
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
Zu, Y., Zheng, Z., Zhu, G., & Jing, Y. P. 2008, ApJ, 686, 41