Galaxy clustering is a standard cosmological probe, which is commonly analysed through two-point statistics. In observations, the estimation of the two-point correlation function crucially relies on counting pairs in a random catalogue. The latter contains a large number of randomly distributed points, which accounts for the survey window function. Random pair counts can also be advantageously used for modelling the window function in the observed power spectrum. Because pair counting scales as $O(N^2)$, where $N$ is the number of points, the computational time to measure random pair counts can be very expensive for large surveys. In this work we present an alternative approach for estimating those counts that does not rely on the use of a random catalogue. We derive an analytical expression for the anisotropic random-random pair counts that accounts for the galaxy radial distance distribution, survey geometry, and possible galaxy weights. We show that a prerequisite is the estimation of the two-point correlation function of the angular selection function, which can be obtained efficiently using pixelated angular maps. Considering the cases of the VIPERS and SDSS-BOSS redshift surveys, we find that the analytical calculation is in excellent agreement with the pair counts obtained from random catalogues.

The main advantage of this approach is that the main calculation only takes a few minutes on a single CPU, and does not depend on the number of random points. Furthermore, it allows an accuracy equivalent to that we would have by using a random catalogue with about 1500 times more points than in the data. We also describe and test an approximate expression for data-random pair counts that is less accurate than for random-random counts, but still provides subpercent accuracy. The presented formalism should be very useful to account for the window function in next-generation surveys, which will necessitate accurate two-point window function estimates over huge observed cosmological volumes.

1. Introduction

The galaxy spatial distribution has a long history of providing cosmological parameter constraints (e.g. Strauss et al. [1992], Vegeley et al. [1992], Maddox et al. [1996], Peacock et al. [2001], Cole et al. [2005], Tegmark et al. [2006], Percival et al. [2010], Blake et al. [2012], de la Torre et al. [2013], Alam et al. [2017], eBOSS Collaboration et al. [2020] and references therein). This arises from the fact that galaxies trace the overall matter distribution, whose statistical properties can be predicted by cosmological models. When analysing galaxy clustering, one usually compresses the information by using summary statistics, the most natural one being the two-point correlation function or its Fourier counterpart the power spectrum. This is due to the nearly Gaussian nature of primordial matter perturbations, which are almost fully described by their two-point statistics. Although gravitational evolution creates non-Gaussianity, and in turn, non-vanishing higher-order $n$-point statistics, two-point statistics remains very informative.

Despite the cosmological principle that implies the correlation function to be isotropic, i.e. to be only a function of the norm of the separation vector, in practice, because of the way the line-of-sight distance is measured in redshift surveys and the presence of peculiar velocities, the observed correlation function becomes anisotropic. These velocities are induced on large scale by the coherent convergence of matter towards overdensities, as part of the general process of structure growth. This anisotropy makes observed galaxy $n$-point statistics sensitive to the strength of gravity acting on the large-scale structure (Kaiser 1987, Guzzo et al. 2008).

Formally, the two-point correlation function is the excess probability of finding a pair of objects at a given distance, with respect to the expectation in a random Poisson distribution of points. In practice, one relies on statistical estimators to measure the correlation function from galaxy survey data. The first estimator was proposed by Peebles & Hauser (1974), and is of the form $\xi_{PH}(s) = DD(s)/RR(s) - 1$, where $DD$ and $RR$ are the normalised number of distinct pairs separated by a vector $s$, in the data and random samples respectively. The latter sample is constructed such that random points follow the same radial and angular selection functions as the data. Other estimators have been later proposed (Hewett 1982, Davis & Peebles 1983, Hamilton 1993) to reduce the estimation variance, notably induced by discreteness and boundary effects. In particular, the Landy & Szalay (1993) minimum-variance estimator was designed such that for any survey geometry its variance is nearly Poisson. This estimator, defined as

$$\xi_{LS}(s) = \frac{DD(s) - 2DR(s) + RR(s)}{RR(s)},$$

makes use of additional data-random pairs $DR$. To estimate the correlation function, one therefore needs to compute the number of pairs as a function of the separation. For the estimator not to be biased and to minimize variance, the random cata-
logue must be much larger than the data catalogue (Landy & Szalay 1993; Keihänen et al. 2019). One usually considers that taking at least about 50 times more random points than objects in the data is enough to avoid introducing additional variance. A problem is that the computational time for direct pair counting scales as $O(N^2)$, with $N$ the number of elements in a given sample. Nonetheless, the complexity can be reduced to $O(N)$ using appropriate algorithms and various efficient codes have been developed implementing those (e.g. Moore et al. 2001; Jarvis et al. 2004; Alonso 2012; Hearin et al. 2017; Marulli et al. 2016; Sinha & Garrison 2020). For the random-random pairs calculation specifically, additional strategies can be used to speed up the computation beyond parallelization, such as splitting the random sample and averaging the counts over subsamples (Keihänen et al. 2019). Still for large surveys, the computational time for estimating the correlation function, especially random-random pairs, can become an issue, particularly in the future for next-generation surveys such as Euclid (Laureijs et al. 2011) or DESI (DESI Collaboration et al. 2016), which would necessitate random samples as large as about $3 \times 10^8$ objects in several redshift bins.

The role of random-random pair counts in the correlation function estimator is to account for the survey selection function, i.e. the effective observed volume and its impact on the data-data pair counts. In Fourier space instead, common estimators for the power spectrum (e.g. Feldman et al. 1994; Yamamoto et al. 2006) provide a direct estimate of the window-convolved power spectrum, and to be able to compare theoretical predictions to observations, one has to convolve the model power spectrum with the survey window function. This convolution is computationally expensive in likelihood analysis, but can be done efficiently by performing a multiplication in configuration space as proposed by Wilson et al. (2017). The latter showed that the window-convolved power spectrum multipole moments $P_L(k)$ can be written as,

$$\hat{P}_L(k) = \mathcal{H} \left[ \sum_{\ell,p,q} \frac{2\ell+1}{2\ell q} \xi(s) N_{RR}(s) \frac{P_{\ell}(s)}{2\pi^2 \Delta s} \right],$$

where $\mathcal{H}$ denotes the Hankel transform, $A_{\ell,p}^q$ are coefficients, $\xi(s)$ is the model correlation function multipole moment, $N$ is a normalisation factor, $R_{RR}(s)$ are the multipole moments of the random-random pair counts, and $\Delta s$ is the bin size in $s$ (Wilson et al. 2017; Beutler et al. 2017).

Random-random pairs counts are a purely geometrical quantity that depends on cosmology only through the radial selection function, which is defined in terms of the radial comoving distance. In the case of a simple geometry, such as a cubical volume with constant number density and periodic boundary conditions, $RR$ pair counts can be predicted from the appropriately normalised ratio between the spherical shell volume at $s$ and the total volume. In the case of a realistic survey geometry, and taking advantage of radial and angular selection functions being usually uncorrelated, Demina et al. (2018) developed a semi-analytical method to compute the $RR$ and $DR$ pair counts along the directions parallel and transverse to the line-of-sight, but still using a random sample to account for angular correlations.

In this paper, we provide general expressions for the anisotropic $RR$ and $DR$ pair counts in the case of a realistic survey geometry, including the cases for the different definitions of the pair line of sight. We apply this formalism to the VIMOS Public Extragalactic Redshift Survey (VIPERS, Guzzo et al. 2014; Garilli et al. 2014) and Sloan Digital Sky Survey Baryon Oscillation Spectroscopic Survey (SDSS-BOSS, Eisenstein et al. 2011; Dawson et al. 2013), and perform an assessment of the accuracy of the method.

This paper is organised as follows. Section 2 presents the formalism for random-random and data-random pair counts. This formalism is applied and its accuracy assessed in Section 3. We conclude in Section 4.

2. Formalism

In this section we provide the analytic formalism for the random-random and data-random pair counts.

2.1. Random-random pairs

In a survey sample where sources are selected in redshift, the number of sources in a given radial distance interval $[r_{\text{min}}, r_{\text{max}}]$ is

$$N(r_{\text{min}}, r_{\text{max}}) = \int_{r_{\text{min}}}^{r_{\text{max}}} n(r) \, dr,$$

with $n(r)$ the number of sources as function of the radial distance $r$ and

$$n(r) = r^2 \bar{n}(r) \int_0^\pi \sin(\theta) \int_0^{2\pi} W(\theta, \varphi) \, d\theta d\varphi,$$

where $W(\theta, \varphi)$ is the survey angular selection function in spherical coordinates. The latter encodes the probability of observing a source at any angular position on the sky and takes values from 0 to 1. $\bar{n}(r)$ is the source number density given by

$$\bar{n}(r) = \begin{cases} \frac{n(r)}{2\pi r^2 W(r)} & \text{for } r < r_{\text{min}}, \\ 0 & \text{for } r_{\text{min}} < r < r_{\text{max}}, \\ 0 & \text{for } r > r_{\text{max}}, \end{cases}$$

with $(W)$ the angular selection function averaged over the full sky. We note that radial weights, such as Feldman et al. 1994 ones, can be included straightforwardly in the $n(r)$, such that this becomes a weighted radial distribution in the equations. The total number of observed sources is therefore

$$N(r_{\text{min}}, r_{\text{max}}) = \int_{r_{\text{min}}}^{r_{\text{max}}} r^2 \bar{n}(r) \int_0^\pi \sin(\theta) \int_0^{2\pi} W(\theta, \varphi) \, d\theta d\varphi d\theta d\varphi.$$ 

In $RR(s)$, we correlate points at two different positions $r_1$ and $r_2$ and it is convenient to write

$$r_2(r_1, s, \mu) = r_1 \sqrt{1 + 2\mu s/r_1^3 + \left(s/r_1^2\right)^2},$$

with $s = r_2 - r_1$, and $\mu = r_1 - s/r_1 s$. $RR$ is obtained by integrating the angular and radial selection functions first over $(r_1, \theta, \varphi)$ and then over the volume defined by the separation $(s, \theta, \varphi)$ as

$$RR(s_{\text{min}}, s_{\text{max}}) = \int_{s_{\text{min}}}^{s_{\text{max}}} \int_{s_{\text{min}}}^{s_{\text{max}}} \int_0^{\pi} \sin \theta d\theta \int_0^{2\pi} d\varphi \bar{n}(r_1) W(\theta_1, \varphi_1) W(\theta_2, \varphi_2) d\theta_1 d\theta_2 d\varphi_1 d\varphi_2,$$

where $\theta_1, \varphi_1, (\theta_2, \varphi_2)$ are the angular positions at $r_1$ ($r_2$). Let us define $n_1 = r_1/r_2$ and $n_2 = r_2/r_1$, we have then

$$\int d^3 n_1 W(n_1) W(n_1 + \phi) = \int d^3 n_1 W(n_1) W(n_1 + \phi).$$
and the correlation function of the angular selection function is
\[ \omega(\phi) = \langle W(n_1)W(n_1 + \phi) \rangle = \frac{1}{4\pi} \int_{4\pi} d\Omega n_1 W(n_1) W(n_1 + \phi). \] (10)

In our case, since we auto-correlate randoms points, RR will only depend on the angular separation (for cross-correlations with different angular selection functions one would need to keep the angular dependence). This means that we can write \( \omega(\phi) = \omega(\phi) \), where we have
\[ \phi(r_1, r_2, s) = \arccos \left( \frac{r_1^2 + r_2^2 - s^2}{2r_1r_2} \right). \] (11)

Note that in the absence of angular mask, i.e. when one evenly probes the full sky, \( \omega(\phi) = 1 \). Putting everything together we find that
\[ RR(s_{\min}, s_{\max}, \mu_{\min}, \mu_{\max}) = 8\pi^2 \int_{\mu_{\min}}^{\mu_{\max}} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} n(r_1) n(r_2) \omega(\phi) dr_1 dr_2 dr_3 d\mu. \] (12)

We note that we have implicitly assumed an end-point definition for the pair line of sight, i.e. for every separation the line-of-sight direction coincides with that of \( r_1 \). With this definition, we can just use \( \omega(\phi) = \omega(\phi) \) for the integral limits in Eq. (12). In the case of the mid-point definition for the pair line of sight, where \( \mu = r \cdot s / r s \) with \( r = \frac{1}{2}(r_1 + r_2) \), we can use the same equation but we need to change the integral limits for each \( r_1, s, \mu \) as
\[ \mu^*(r_1, s, \mu) = -s + s \mu^2 + \mu \sqrt{s^2 \mu^2 - s^2 + 4r_1^2}. \] (13)

We note that for the end-point definition it is important to compute pairs with \( \mu < 0 \) since the correlation function in that case is not symmetric by pair exchange. For applications where one is interested in the random-random multipole moments directly, the latter can be defined as
\[ RR(s_{\min}, s_{\max}) = 8\pi^2 \int_{\mu_{\min}}^{\mu_{\max}} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} n(r_1) n(r_2) \omega(\phi) L_\ell(\mu^*) dr_1 dr_2 dr_3 d\mu. \] (14)

with \( L_\ell \) the Legendre polynomial of order \( \ell \). For the end-point definition \( \mu^* = \mu \), while for the mid-point definition we have
\[ \mu^*(r_1, s, \mu) = \frac{r_1}{r} + \frac{1}{2} s, \] (15)

with \( r(r_1, s, \mu) = r_1 \sqrt{1 + \mu s / r + s^2 / (2r)} \). Finally, we note that in order to cross-correlate tracers with different radial selection functions but the same angular selection function, we can use the same formalism but with different \( n_A(r_1) \) and \( n_B(r_2) \) in Eq. (12) (14).

2.2. Data-random pairs

The [Landy & Szalay (1993)] estimator includes data-random pairs to minimise variance. A similar formalism to that used for random-random pair counts can be employed to evaluate data-random pair counts. But contrarily to the RR case, we now have to cross-correlate a discrete set of sources with a continuous random distribution. The discrete limit of Eq. (15) is
\[ N(r_{\min}, r_{\max}) = \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} \delta_D(r - r_0) d\Omega d^2 n. \] (16)

with \( \delta_D \) the Dirac delta function, \( r = (r, \theta, \varphi) \), and \( r_0 \) the sources position in the data vector. We can then use the same methodology as in Section 2.1. To make the computation of the data-random pair counts tractable we make the assumption
\[ \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} d^2 n_1 \delta_D(r_1 - r_0) W_n(1 + \phi) = \frac{1}{N} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} d^2 n_1 \delta_D(n_1 - n_0) W(n_1 + \phi). \] (17)

with \( n_1 = (\theta, \varphi) \) and \( n_0 \) the angular position in the data vector. Under this assumption, the angular correlation function at some point is given by that of the whole sample. For a large enough \( N \) and if the angular sampling of the data is sufficiently homogeneous, the approximation should hold. We find that
\[ DR(s_{\min}, s_{\max}, \mu_{\min}, \mu_{\max}) = \frac{4\pi}{N} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} \delta_D(r_1 - r_0) \int_{\mu_{\min}}^{\mu_{\max}} \int_{r_{\min}}^{r_{\max}} n(r_2) \int_{0}^{2\pi} \omega_{DR}(\phi) d\Omega d\mu d\varphi. \] (18)

where \( \omega_{DR}(\phi) = \frac{1}{4\pi} \int_{4\pi} d\Omega n_1 \delta_D(n_1 - n_0) W(n_1 + \phi) \). (19)

If we further assume that the angular cross-correlation function does not depend on the pair orientation, we obtain
\[ DR(s_{\min}, s_{\max}, \mu_{\min}, \mu_{\max}) = \frac{8\pi^2}{N} \sum_{r_1 \in [r_{\min}, r_{\max}]} w_i \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r_{\max}} n(r_2) \omega_{DR}(\phi) d\Omega d\mu d\varphi. \] (20)

We can already see that the calculation involves a sum over all data sources, which is potentially more computationally expensive than in the RR and direct pair-counting cases. Nonetheless, to make the computation efficient we can approximate this by taking the continuous limit on the sum and write similarly as for the RR case
\[ DR(s_{\min}, s_{\max}, \mu_{\min}, \mu_{\max}) = 8\pi^2 \int_{r_{\min}}^{r_{\max}} r_1^2 n_1(r_1) \int_{r_{\min}}^{r_{\max}} s^2 \int_{r_{\min}}^{r_{\max}} n_2(r_2) \omega_{DR}(\phi) dr_1 dr_2 d\mu. \] (21)

We can already see that the calculation involves a sum over all data sources, which is potentially more computationally expensive than in the RR and direct pair-counting cases. Nonetheless, to make the computation efficient we can approximate this by taking the continuous limit on the sum and write similarly as for the RR case
\[ DR(s_{\min}, s_{\max}, \mu_{\min}, \mu_{\max}) = 8\pi^2 \int_{r_{\min}}^{r_{\max}} r_1^2 n_1(r_1) \int_{r_{\min}}^{r_{\max}} s^2 \int_{r_{\min}}^{r_{\max}} n_2(r_2) \omega_{DR}(\phi) dr_1 dr_2 d\mu. \] (22)
where the angular cross-correlation function $\omega_{DR}$ can be written as $\omega_{DR} = (W_1 W_2)$, with $W_1$ a pixelated map containing (weighted) source counts and $W_2$ the angular selection function as in Section 2.1. In this case, the normalisation on $W_1$ and $W_2$ does not matter since the final result is proportional to $\omega_{DR}/W_1 W_2$. In principle, Eq. (22) should be close to Eq. (21) when the $n(r)$ is fine enough to faithfully reproduce data radial overdensities.

3. Application

In this section we apply our formalism for RR and DR pair counts to the case of BOSS and VIPERS redshift surveys, and test its accuracy.

3.1. Numerical implementation

The method takes as input the radial distance distribution of sources and the correlation function of the angular selection function. To compute the latter, we first produce a Healpix (Gorski et al. 2005) full-sky map from the survey angular mask, which is generally in form of a list of distinct spherical polygons with associated weights. We infer the angular correlation function from the maps using PolSpice (Szapudi et al. 2001; Chon et al. 2004), which takes advantage of Healpix isolatitude pixelation scheme to fast evaluate the angular correlation function $\omega(\theta)$ with fast spherical harmonic transforms. The CPU time to compute the angular correlation function depends on the map resolution, but is in general quite efficient. The map resolution is controlled by the $nside$ parameter. For instance, for BOSS it takes 3, 20 and 130 minutes on a single CPU for $nside = 2048$, 4096, and 8192 while for VIPERS it takes 6, 40 and 320 minutes for $nside = 8192$, 16384, and 32768. Note that we compute the $\omega_{DR}$ with associated weights. We infer the angular correlation function $\langle \omega \rangle = \langle W \rangle \langle W \rangle^*$, which is generally in form of a list of distinct spherical polygons.

In each case, we need to specify a maximum tolerance on the integral relative error $\epsilon$. The GSL cquad algorithm only performs one-dimensional integrals, and we thus implement nested integrals with same $\epsilon$ to perform the full three-dimensional integral. There are three potential sources of error associated to the analytical pair counts calculation: the $n(r)$ estimation, the $\omega(\theta)$ estimation, and the precision on numerical integrals. In our methodology, the error level associated to each source is controlled respectively by the binning in $n(r)$, angular map resolution $nside$, and $\epsilon$. In the last case however, we note that different algorithms can yield slightly different results even if $\epsilon$ is very small.

Regarding the performances of the RR implementation, for the two considered surveys and considering 6000 bins in $(s,\mu)$, the full $RR(s,\mu)$ based on Eq. (12) takes about 5-20 minutes on a single CPU using the CUBA library, even for $\epsilon \approx 10^{-6}$ (GSL takes significantly more time when $\epsilon < 10^{-2}$). In principle, one gains an additional factor of two when using the pair line-of-sight mid-point definition, as in that case, there is a symmetry along the line of sight for auto-correlation and one only needs to compute $\mu > 0$ pairs. The run time depends on the number of bins in $(s,\mu)$ but also in principle, on the shape of the integrand, as for complex $n(r)$ or $\omega(\theta)$ the integrals will take more time to converge (although in our case we did not see any noticeable difference). For the RR multipole moments in Eq. (14), the calculation only takes about 5 seconds with CUBA algorithms using 30 bins in $s$, independently of the adopted value of $\epsilon$.

Regarding DR, the run times for the approximation in Eq. (22) are similar to those for RR by definition. However, the evaluation of Eq. (21) leads to large computational times of up to several weeks on a single CPU for large datasets. The run times scale linearly with the number of objects in the data sample. In that case, direct data-random pair counting might be more efficient.

A C code that follows this implementation is publicly available at http://github.com/mianbreton/RR_code. It can be used with any input $n(r)$ and $\omega(\theta)$ to predict RR or DR pair counts.

3.2. Survey selection functions

We consider two realistic redshift survey selection functions, those of SDSS-BOSS DR12 CMASS (Alam et al. 2015) and VIPERS PDR2 (Scodeggio et al. 2013) galaxy samples. We use the public galaxy catalogue and associated angular masks. Those two samples have complementary properties and thus allow testing the method in different conditions. Indeed, while BOSS survey is wide and has a low galaxy number density, VIPERS is much narrower and denser. Each survey is composed of two separated fields on the sky but we only consider here BOSS North Galactic Cap (NGC) and VIPERS W1 fields, and we focus on $0.5 < z < 0.75$ and $0.7 < z < 1.2$ redshift intervals for BOSS and VIPERS respectively.

The radial selection functions are shown in Fig 1. The BOSS $n(r)$ is estimated from the data by taking the histogram of galaxy comoving distances and cubic-spline interpolating between the bins. In the case of VIPERS, we use the fitting function for the radial distribution given in de la Torre et al. (2013). We assumed two cosmologies to convert redshift to comoving distance: flat $\Lambda$CDM with $\Omega_m = 0.31$ and $\Omega_m = 0.25$ for BOSS and VIPERS respectively. Nonetheless, the choice of fiducial cosmology has no impact on the accuracy of the analytical predictions.

The angular selection functions that we used for VIPERS W1 and SDSS-BOSS CMASS NGC are presented in Appendix A. They enter in Eq. (12) through their auto-correlation function. The latter are given in Fig 2 and Fig 3, respectively for BOSS and VIPERS. We test different map resolutions by varying the Healpix resolution parameter $nside$ from 2048 to 8192. For BOSS, the correlation function is very smooth. The relative difference between $nside = 2048$, 4096 cases and $nside = 8192$ is roughly constant, at 0.1% and 0.05% respectively. In the case of VIPERS, the angular mask has more small-scale features but similarly, the relative differences between angular correlation functions based on different map resolutions are nearly constant in scale. The bias is larger than in the BOSS case, with a rel-

1 The total number of pixels in a full-sky map is given by $N_{pix} = 12 \times nside^2$.

2 Available at http://data.sdss.org/sas/drl/BOSS/DR12/BOSS/lss/BOSS and http://vipers.inaf.it/rel-pdr2.html (VIPERS).
Fig. 1. Adopted radial distance distribution $n(r)$ for the BOSS NGC CMASS sample at $0.5 < z < 0.75$ (blue solid curve) and VIPERS W1 sample at $0.7 < z < 1.2$ (red dashed line). The distributions are normalised so that the integral is unity. The vertical solid (dashed) lines show the adopted sample limits for BOSS (VIPERS).

Fig. 2. Top panel: two-point correlation function of the BOSS angular selection function obtained from a Healpix map with $nside = 8192$. Bottom panel: relative difference on the angular two-point correlation function with respect to lower resolution maps, i.e. $nside = 2048, 4096$ in blue and orange curves respectively.

Fig. 3. Top panel: two-point correlation function of the VIPERS angular selection function obtained from a Healpix map with $nside = 65536$. Bottom panel: relative difference on the angular two-point correlation function with respect to lower resolution maps, i.e. $nside = 8192, 16834, 32768$ in blue, orange, and green curves respectively.

Fig. 4. Detail of the VIPERS W1 angular mask showing the impact of Healpix resolution on the sampling of the survey angular mask.

3.3. $RR$ counts

We compare our analytical prediction for $RR$ with the average random-random counts ($RR$), obtained from 100 random samples constructed using the same radial and angular selection functions. Within the considered redshift intervals, there are 435185 BOSS and 24316 VIPERS galaxies and we use $3 \times 10^7$ and $3.9 \times 10^6$ points per random sample respectively (i.e. multiplicative factors of about 70 and 160 with respect to the data). We compute the pair counts from the random samples using the fast Corrfunc pair-counting code (Sinha & Garrison 2020). Our method predicts anisotropic $RR(s, \mu)$ counts, but to simplify the comparison, we simply consider the monopole, i.e. $RR_0(s) = \sum_i RR(s, \mu_i) \Delta \mu$, where linear bins $\mu_i$ extend from $-1$ to $1$. Those comparisons are presented in Fig. 5 for BOSS and
in Fig. 5 for VIPERS. We see that for both surveys, the relative difference between the analytical computation and \(\langle RR\rangle\) is well within the variance of the random samples. We compare the results obtained with different numerical integration algorithms (see figure insets and Section 3.1) and find that \textit{cuhre} tends to depart from the others algorithms, which is understandable since it is intrinsically different from the others. If we ignore \textit{cuhre}, we see that at most the relative difference between the analytical computation and \(\langle RR\rangle\) remains within \(3 \times 10^{-3}\) for BOSS and \(1.7 \times 10^{-4}\) for VIPERS.

The variance on the random sample counts depends on the number of points in the sample, and one can ask what is the number of random points needed to achieve the same accuracy as in the analytical method. Keihänen et al. (2019) showed that the relative variance on \(RR\) in a given bin is

\[
\text{var}(RR) = \frac{2}{N_r(N_r - 1)} \left\{ \left(2N_r - 1\right) \left( \frac{G'}{\langle G'\rangle^2} - 1 \right) + \frac{1}{\langle G'\rangle^2} - 1 \right\} \tag{23}
\]

with \(N_r\) the number of random points, and \(G', G''\) terms are (Landy & Szalay 1993)

\[
G' = \frac{\langle n_r \rangle}{N_r(N_r - 1)/2}, \tag{24}
\]

\[
G'' = \frac{\langle n_r \rangle}{N_r(N_r - 1)(N_r - 2)/2}, \tag{25}
\]

with \(\langle n_r \rangle\) and \(\langle n_t \rangle\) the number of pairs and triplets averaged over several realisations. While \(G'\) can easily be estimated from the random samples, we directly solve for \(G''\) from the estimated \text{var}(RR). We can then deduce which \(N_r\) give standard deviations similar to \(3 \times 10^{-3}\) and \(1.7 \times 10^{-4}\). We found that we need an additional factor of \textit{at least} 20 (10) for BOSS (VIPERS) in the number of random points. Therefore, the analytical method allows the achievement of the same accuracy as by using a random sample with about \(20 \times 70 (10 \times 160)\) more points than data in BOSS (VIPERS). We finally note that CUBA integration algorithms have parameters that can be potentially further fine-tuned to achieve better accuracy.

### 3.4. \textit{DR} counts

In the \textit{DR} case, we need to rely on approximations. Under the approximation in Eq. (17), we have two possibilities to calculate \textit{DR} counts: either a discrete sum over all source distances as in Eq. (21) or by further approximating the discrete sum by an integral as in Eq. (22). In the last case, we can already anticipate that the results will depend on the input \(\bar{n}(r_1)\), particularly its ability to reproduce line-of-sight structures in the data. In Fig. 7 and Fig. 8 we show different estimations of the data \(\bar{n}(r_1)\) in VIPERS and BOSS, varying the bin size in \(r_1\). In the limit where \(\bar{n}(r_1)\) resembles a sum of Dirac delta functions, Eq. (22) should be equivalent to Eq. (21). For the random part we use in \(\bar{n}(r_2)\) the distributions provided in Fig. 1.

Following the same methodology as in Section 3.3 we compute \(\langle DR\rangle\) for both surveys using the same data catalogue and 100 random samples, which we later compare to the predictions based on Eq. (21) and Eq. (22) using different input data \(\bar{n}(r_1)\).
Michel-Andrés Breton and Sylvain de la Torre: Fast RR and DR calculation

Fig. 8. Same as Fig. 7 but for BOSS.

Fig. 9. Relative difference between the analytical and random catalogue-based mean $\langle DR \rangle$ pair counts monopole for VIPERS. The grey shaded area shows the standard deviation among the random catalogues. The blue, red, and black predictions use Eq. (17) and $\bar{n}(r)$ with large (blue), intermediate (red), and small (black) bin sizes. These use vegas algorithm with $\epsilon = 10^{-5}$. The green line shows the prediction of Eq. (21) obtained with GSL using $\epsilon = 10^{-4}$.

In the case of VIPERS, we find that when using Eq. (21) the discrepancy between the analytical prediction and direct pair counting is of the order of 1%, as shown in Fig. 9. Moreover, we see that the prescription in Eq. (22) leads to a systematic bias of up to about 2% when using a large binning in the input $\bar{n}(r_1)$, but converges towards Eq. (21) result when a small binning is adopted, as expected.

In the case of BOSS, we find similar trends but with an higher accuracy, as shown in Fig. 10. We find at most a difference of $5 \times 10^{-4}$ between the analytical solution, either Eq. (21) or Eq. (22) with a fine $\bar{n}(r_1)$, and direct pair counting. Here the approximation in Eq. (17) is more appropriate since the data sample is larger. This explains the better reached accuracy. We emphasize that the variance in Fig. 9 and Fig. 10 only comes from the random samples, since a single data catalogue is used. Therefore, increasing the number of random points would reduce this variance. If we were to correlate different data samples with those random samples in order to obtain the variance on $DR$, the the latter would be much larger than in the present case. Overall, because of the approximation in Eq. (17), our analytical $DR$ predictions remain biased, exceeding the typical variance introduced by random sampling in direct pair counting.

4. Conclusion

In this paper, we presented general analytical expressions for the random-random and data-random pair counts, in the case of a realistic survey geometry. The main results are given in Eq. (12) (or Eq. 14 for the multipole moments) for $RR$ and in Eq. (22) for $DR$. These expressions can be solved numerically in an efficient way. This method, which does not rely on generating random mocks, only takes as input the comoving radial distance distribution in an assumed cosmology, and the angular selection function two-point correlation function, which only needs to be estimated once for a given survey. Once those quantities are provided, the full computation takes about a few minutes to obtain anisotropic pair counts $RR(s, \mu)$ and a few seconds for its multipole moments, using a single CPU and standard libraries for three-dimensional integration.

We tested this method in the context of the BOSS and VIPERS survey geometries, and found excellent agreements with expected $RR$ pair counts. The predicted counts exhibit a high accuracy, equivalent to that we would obtain by performing pair counting in random samples of about 1400-1600 more random points than data in those surveys. The main advantage is that the method is fast and does not rely on any spatial sampling, while usually one needs to generate a random catalogue with at least 50 times the number of objects in the data. We believe that this can be of some use for future surveys with large data sample and very expensive $RR$ pair counts calculation.

The $DR$ pair counts can also be calculated analytically under some approximations. We found that the results are slightly biased with respect to the expected counts. For VIPERS and BOSS we found a bias with respect to direct pair counts of 1% and 0.05% respectively. This bias should decrease with the increasing number of data points. When estimating $DR$ for several data samples, one needs to compute for each sample its angular two-point correlation function with respect to the survey angular selection function.

Article number, page 7 of 9
Overall, the method presented in this paper to efficiently evaluate the survey window two-point function should be very useful to deal with massive galaxy surveys. The provided formulae are fast to evaluate, and with further efficient parallelization [Hahn2015], e.g., one should be able to compute \( RR \) and \( DR \) in an extremely small amount of time. In that case, one could imagine \( RR \) and \( DR \) being evaluated in different cosmologies at each step of a cosmological likelihood analysis. This opens new horizons in the way we should analyse galaxy survey data in the future.

Acknowledgements. We thank Eric Jullo for his help on dealing with partial, high-resolution Healpix maps and his comments on the draft. We thank the Instituto de Astrofísica de Andalucía (IAA-CSIC), and the Spanish academic and research network RedIRIS [http://www.rediris.es] in Spain for providing the skun at IAA, RedIRIS server that allowed us to run the calculations for high-resolution \( \text{nside} = 65535 \) maps. This research used resources of the Dark Energy computing Center funded by the OCEVU Labex (ANR-11-LABX-0060). Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is [http://www.sdss3.org](http://www.sdss3.org).

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

References

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Alam, S., Ata, M., Bailey, S., et al. 2017, MNRAS, 470, 2617
Alonso, D. 2012, ArXiv e-prints [arXiv:1210.1833]
Beutler, F., Sco, H.-J., Saito, S., et al. 2017, MNRAS, 466, 2242
Blake, C., Brough, S., Colless, M., et al. 2012, MNRAS, 425, 405
Chon, G., Challinor, A., Prunet, S., Hivon, E., & Szapudi, I. 2004, MNRAS, 350, 914
Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505
Davis, M. & Peebles, P. J. E. 1983, ApJ, 267, 465
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
de la Torre, S., Guzzo, L., Peacock, J. A., et al. 2013, A&A, 557, A54
Demina, R., Cheong, S., BenZvi, S., & Hindrichs, O. 2018, MNRAS, 480, 49
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv e-prints, arXiv:1611.00036
eBOSS Collaboration, Alam, S., Aubert, M., et al. 2020, arXiv e-prints, arXiv:2007.08991
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, ApJ, 426, 23
Garilli, B., Guzzo, L., Scoccoeggi, M., et al. 2014, A&A, 562, A23
Görski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Gough, B. 2009, GNU scientific library reference manual (Network Theory Ltd.)
Guzzo, L., Pierleoni, M., Meneux, B., et al. 2008, Nature, 451, 541
Guzzo, L., Scoccoeggi, M., Garilli, B., et al. 2014, A&A, 566, A108
Hahn, T. 2008, Computer Communications, 168, 78
Hahn, T. 2015, in Journal of Physics Conference Series, Vol. 608, Journal of Physics Conference Series, 012066
Hamilton, A. J. S. 1993, ApJ, 417, 19
Hearin, A. P., Campbell, D., Tollerud, E., et al. 2017, AJ, 154, 190
Hewett, P. C. 1982, MNRAS, 201, 867
Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338
Kaiser, N. 1987, MNRAS, 227, 1
Keihänen, E., Kurki-Suonio, H., Lindholm, V., et al. 2019, A&A, 631, A73
Landy, S. D. & Szalay, A. S. 1993, ApJ, 412, 64
Lau, L., Amaux, J., Ardiani, S., et al. 2011, arXiv e-prints, arXiv:1110.3193
Maddox, S. J., Efstathiou, G., & Sutherland, W. J. 1996, MNRAS, 283, 1227
Marulli, F., Veropalumbo, A., & Moresco, M. 2016, Astronomy and Computing, 14, 35
Moore, A. W., Connolly, A. J., Genovese, C., et al. 2001, in Mining the Sky, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann, 71
Peacock, J. A., Cole, S., Norberg, P., et al. 2001, Nature, 410, 169

Article number, page 9 of 9

Peebles, P. J. E. & Hauser, M. G. 1974, APJS, 28, 19
Percival, W. J., Reid, B. A., Eisenstein, D. J., et al. 2010, MNRAS, 401, 2148
Scoccoeggi, M., Guzzo, L., Garilli, B., et al. 2018, A&A, 609, A84
Sinha, M. & Garrison, L. H. 2020, MNRAS, 491, 3022
Strauss, M. A., Davis, M., Yahil, A., & Huchra, J. P. 1992, ApJ, 385, 421
Szapudi, I., Prunet, S., Pogosyan, D., Szalay, A. S., & Bond, J. R. 2001, ApJ, 548, L115
Tegmark, M., Eisenstein, D. J., Strauss, M. A., et al. 2006, Phys. Rev. D, 74, 123507
Vogeley, M. S., Park, C., Geller, M. J., & Huchra, J. P. 1992, ApJ, 391, L5
Wilson, M. J., Peacock, J. A., Taylor, A. N., & de la Torre, S. 2017, MNRAS, 464, 3121
Yamamoto, K., Nakamichi, M., Camino, A., Basset, B. A., & Nishioka, H. 2006, PASJ, 58, 93
Appendix A: VIPERS and SDSS-BOSS survey footprints

In Figs. A.1 and A.2 we provide the footprints and angular masks for VIPERS W1 and SDSS-BOSS CMASS NGC fields respectively, which we used in this analysis. In the case of BOSS angular mask, each distinct mask polygon has an associated tiling success rate, which is a measure of the completeness in associating fibres to potential spectroscopic targets in the survey. We use this quantity as a weight in defining the angular selection function. In the case of VIPERS, the angular selection function is taken to be unity inside the spectroscopic mask (quadrant-shaped polygons) and null otherwise, except in the regions of the photometric mask (circular- and star-shaped polygons) where it is also set to zero.

Fig. A.1. VIPERS W1 footprint.

Fig. A.2. BOSS CMASS NGC footprint.