HAL
open science

# Estimating the galaxy two-point correlation function using a split random catalog

E. Keihänen, H. Kurki-Suonio, V. Lindholm, A. Viitanen, A.-S. Suur-Uski, V. Allevato, E. Branchini, F. Marulli, P. Norberg, D. Tavagnacco, et al.

## To cite this version:

E. Keihänen, H. Kurki-Suonio, V. Lindholm, A. Viitanen, A.-S. Suur-Uski, et al.. Estimating the galaxy two-point correlation function using a split random catalog. Astronomy and Astrophysics - A&A, 2019, 631, pp.A73. 10.1051/0004-6361/201935828 . cea-02327633

**HAL Id: cea-02327633**

**https://cea.hal.science/cea-02327633v1**

Submitted on 22 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A&A 631, A73 (2019)
https://doi.org/10.1051/0004-6361/201935828
© ESO 2019

Astronomy & Astrophysics

# Estimating the galaxy two-point correlation function using a split random catalog

E. Keihänen[1], H. Kurki-Suonio[1], V. Lindholm[1], A. Viitanen[1], A.-S. Suur-Uski[1], V. Allevato[2,1], E. Branchini[3], F. Marulli[4,5,6], P. Norberg[7], D. Tavagnacco[8], S. de la Torre[9], J. Valiviita[1], M. Viel[10,11,8,14], J. Bel[12], M. Frailis[8], and A. G. Sánchez[13]

[1] Department of Physics and Helsinki Institute of Physics, University of Helsinki, Gustaf Hällströmin katu 2, 00014 Helsinki, Finland
e-mail: elina.keihanen@helsinki.fi
[2] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
[3] Department of Mathematics and Physics, Roma Tre University, Via della Vasca Navale 84, 00146 Rome, Italy
[4] Dipartimento di Fisica e Astronomia – Alma Mater Studiorum Università di Bologna, Via Piero Gobetti 93/2, 40129 Bologna, Italy
[5] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
[6] INFN – Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[7] ICC & CEA, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK
[8] INAF, Osservatorio Astronomico di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
[9] Aix Marseille Univ., CNRS, CNES, LAM, Marseille, France
[10] SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, TS, Italy
[11] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste, TS, Italy
[12] Aix Marseille Univ., Université de Toulon, CNRS, CPT, Marseille, France
[13] Max-Planck-Institut für Extraterrestrische Physik, Postfach 1312, Giessenbachstr., 85741 Garching, Germany
[14] IFPU – Institute for Fundamental Physics of the Universe, Via Beirut 2, 34014 Trieste, Italy

Received 3 May 2019 / Accepted 24 July 2019

**ABSTRACT**

The two-point correlation function of the galaxy distribution is a key cosmological observable that allows us to constrain the dynamical and geometrical state of our Universe. To measure the correlation function we need to know both the galaxy positions and the expected galaxy density field. The expected field is commonly specified using a Monte-Carlo sampling of the volume covered by the survey and, to minimize additional sampling errors, this random catalog has to be much larger than the data catalog. Correlation function estimators compare data–data pair counts to data–random and random–random pair counts, where random–random pairs usually dominate the computational cost. Future redshift surveys will deliver spectroscopic catalogs of tens of millions of galaxies. Given the large number of random objects required to guarantee sub-percent accuracy, it is of paramount importance to improve the efficiency of the algorithm without degrading its precision. We show both analytically and numerically that splitting the random catalog into a number of subcatalogs of the same size as the data catalog when calculating random–random pairs and excluding pairs across different subcatalogs provides the optimal error at fixed computational cost. For a random catalog fifty times larger than the data catalog, this reduces the computation time by a factor of more than ten without affecting estimator variance or bias.

**Key words.** large-scale structure of Universe – cosmology: observations – methods: statistical – methods: data analysis

## 1. Introduction

The spatial distribution of luminous matter in the Universe is a key diagnostic for studying cosmological models and the physical processes involved in the assembly of structure. In particular, light from galaxies is a robust tracer of the overall matter distribution, whose statistical properties can be predicted by cosmological models. Two-point correlation statistics are very effective tools for compressing the cosmological information encoded in the spatial distribution of the mass in the Universe. In particular, the two-point correlation function in configuration space has emerged as one of the most popular cosmological probes. Its success stems from the presence of characterized features that can be identified, measured, and effectively compared to theoretical models to extract clean cosmological information.

One such feature is baryon acoustic oscillations (BAOs), which imprint a characteristic scale in the two-point correlation function that can be used as a standard ruler. After the first detection in the two-point correlation function of SDSS DR3 and 2dFGRS galaxy catalogs (Eisenstein et al. 2005; Cole et al. 2005), the BAO signal was identified, with different degrees of statistical significance, and has since been used to constrain the expansion history of the Universe in many spectroscopic galaxy samples (see e.g., Percival et al. 2010; Blake et al. 2011; Beutler et al. 2011; Anderson et al. 2012, 2014; Ross et al. 2015, 2017; Alam et al. 2017; Vargas-Magaña et al. 2018; Bautista et al. 2018; Ata et al. 2018). Several of these studies did not focus on the BAO feature only but also analyzed the anisotropies in the two-point correlation function induced by the peculiar velocities (Kaiser 1987), the so-called redshift space distortions (RSD), and by assigning cosmology-dependent distances to the observed redshifts (the Alcock & Paczyński 1979 test). For RSD analyses, see also, for example, Peacock et al. (2001), Guzzo et al. (2008), Beutler et al. (2012), Reid et al. (2012), de la Torre et al. (2017),

Pezzotta et al. (2017), Zarrouk et al. (2018), Hou et al. (2018), and Ruggeri et al. (2019).

Methods to estimate the galaxy two-point correlation function (2PCF) $\xi(\boldsymbol{r})$ from survey data are based on its definition as the excess probability of finding a galaxy pair. One counts from the data ($D$) catalog the number $DD(\boldsymbol{r})$ of pairs of galaxies with separation $\boldsymbol{x}_2 - \boldsymbol{x}_1 \in \boldsymbol{r}$, where $\boldsymbol{r}$ is a bin of separation vectors, and compares it to the number of pairs $RR(\boldsymbol{r})$ in a corresponding randomly generated ($R$) catalog and to the number of data-random pairs $DR(\boldsymbol{r})$. The bin may be a 1D ($r \pm \frac{1}{2}\Delta r$), 2D, or a 3D bin. In the 1D case, $r$ is the length of the separation vector and $\Delta r$ is the width of the bin. From here on, "separation $\boldsymbol{r}$" indicates that the separation falls in this bin.

Several estimators of the 2PCF have been proposed by Hewett (1982), Davis & Peebles (1983), Hamilton (1993), and Landy & Szalay (1993), building on the original Peebles & Hauser (1974) proposal. These correspond to different combinations of the $DD$, $DR$, and $RR$ counts to obtain a 2PCF estimate $\hat{\xi}(\boldsymbol{r})$; see Kerscher (1999) and Kerscher et al. (2000) for more estimators. The Landy–Szalay (Landy & Szalay 1993) estimator

$$\hat{\xi}_{\rm LS}(\boldsymbol{r}) := \frac{N_{\rm r}'(N_{\rm r}'-1)}{N_{\rm d}(N_{\rm d}-1)}\frac{DD(\boldsymbol{r})}{RR(\boldsymbol{r})} - \frac{N_{\rm r}'-1}{N_{\rm d}}\frac{DR(\boldsymbol{r})}{RR(\boldsymbol{r})} + 1, \tag{1}$$

(we call this method "standard LS" in the following) is the most commonly used, since it provides the minimum variance when $|\xi| \ll 1$ and is unbiased in the limit $N_{\rm r}' \to \infty$. Here $N_{\rm d}$ is the size (number of objects) of the data catalog and $N_{\rm r}'$ is the size of the random catalog. We define $M_{\rm r} := N_{\rm r}'/N_{\rm d}$. To minimize random error from the random catalog, $M_{\rm r} \gg 1$ should be used (for a different approach, see Demina et al. 2018).

One is usually interested in $\xi(\boldsymbol{r})$ only up to some $r_{\rm max} \ll L_{\rm max}$ (the maximum separation in the survey), and therefore pairs with larger separations can be skipped. Efficient implementations of the LS estimator involve pre-ordering of the catalogs through kd-tree, chain-mesh, or other algorithms (e.g., Moore et al. 2000; Alonso 2012; Jarvis 2015; Marulli et al. 2016) to facilitate this. The computational cost is then roughly proportional to the actual number of pairs with separation $r \le r_{\rm max}$.

The correlation function is small for large separations, and in cosmological surveys $r_{\rm max}$ is large enough so that for most pairs $|\xi(\boldsymbol{r})| \ll 1$. The fraction $f$ of $DD$ pairs with $r \le r_{\rm max}$ is therefore not very different from the fraction of $DR$ or $RR$ pairs with $r \le r_{\rm max}$. The computational cost is dominated by the part proportional to the total number of pairs needed, $\frac{1}{2}fN_{\rm d}(N_{\rm d}-1) + fN_{\rm d}N_{\rm r} + \frac{1}{2}fN_{\rm r}'(N_{\rm r}'-1) \approx \frac{1}{2}fN_{\rm d}^2(1+2M_{\rm r}+M_{\rm r}^2)$, which in turn is dominated by the $RR$ pairs as $M_{\rm r} \gg 1$. The smaller number of $DR$ pairs contribute much more to the error of the estimate than the large number of $RR$ pairs, whereas the cost is dominated by $RR$. Thus, a significant saving of computation time with an insignificant loss of accuracy may be achieved by counting only a subset of $RR$ pairs, while still counting the full set (up to $r_{\rm max}$) of $DR$ pairs.

A good way to achieve this is to use many small (i.e, low-density) $R$ catalogs instead of one large (high-density) catalog (Landy & Szalay 1993; Wall & Jenkins 2012; Slepian & Eisenstein 2015), or, equivalently, to split an already generated large $R$ catalog into $M_{\rm s}$ small ones for the calculation of $RR$ pairs while using the full $R$ catalog for the $DR$ counts. This method has been used by some practitioners (e.g., Zehavi et al. 2011[1]), but this is usually not documented in the literature. One might also consider obtaining a similar cost saving by diluting (subsampling) the $R$ catalog for $RR$ counts, but, as we show below, this is not a good idea. We refer to these two cost-saving methods as "split" and "dilution".

In this work we theoretically derive the additional covariance and bias due to the size and treatment of the $R$ catalog; test these predictions numerically with mock catalogs representative of next-generation datasets, such as the spectroscopic galaxy samples that will be obtained by the future Euclid satellite mission (Laureijs et al. 2011); and show that the "split" method, while reducing the computational cost by a large factor, retains the advantages of the LS estimator.

We follow the approach of Landy & Szalay (1993; hereafter, LS93), but generalize it in a number of ways: In particular, since we focus on the effect of the random catalog, we do not work in the limit $M_{\rm r} \to \infty$. Also, we calculate covariances, not just variances, and make fewer approximations (see Sect. 2.2).

The layout of the paper is as follows. In Sect. 2 we derive theoretical results for bias and covariance. In Sect. 3 we focus on the split LS estimator and its optimization. In Sect. 4 we test the different estimators with mock catalogs. Finally, we discuss the results and present our conclusions in Sect. 5.

## 2. Theoretical results: bias and covariance

### 2.1. General derivation

We follow the derivation and notations in LS93 but extend to the case that includes random counts covariance. We consider the survey volume as divided into $K$ microcells (very small subvolumes) and work in the limit $K \to \infty$, which means that no two objects will ever be located within the same microcell.

Here, $\alpha$, $\beta$, and $\gamma$ represent the relative deviation of the $DD(\boldsymbol{r})$, $DR(\boldsymbol{r})$, and $RR(\boldsymbol{r})$ counts from their expectation values (mean values over an infinite number of independent realizations):

$$\begin{aligned} DD(\boldsymbol{r}) &=: \langle DD(\boldsymbol{r})\rangle[1+\alpha(\boldsymbol{r})],\\ DR(\boldsymbol{r}) &=: \langle DR(\boldsymbol{r})\rangle[1+\beta(\boldsymbol{r})],\\ RR(\boldsymbol{r}) &=: \langle RR(\boldsymbol{r})\rangle[1+\gamma(\boldsymbol{r})]. \end{aligned} \tag{2}$$

By definition $\langle\alpha\rangle = \langle\beta\rangle = \langle\gamma\rangle = 0$. The factors $\alpha, \beta$, and $\gamma$ represent fluctuations in the pair counts, which arise as a result of a Poisson process. As long as the mean pair counts per bin are large ($\gg$1) the relative fluctuations will be small. We calculate up to second order in $\alpha, \beta$, and $\gamma$, and ignore the higher-order terms (in the limit $M_{\rm r} \to \infty$, $\gamma \to 0$, so LS93 set $\gamma = 0$ at this point).

The expectation values for the pair counts are:

$$\begin{aligned} \langle DD(\boldsymbol{r})\rangle &= \frac{1}{2}N_{\rm d}(N_{\rm d}-1)G^{\rm p}(\boldsymbol{r})[1+\xi(\boldsymbol{r})],\\ \langle DR(\boldsymbol{r})\rangle &= N_{\rm d}N_{\rm r}G^{\rm p}(\boldsymbol{r}),\\ \langle RR(\boldsymbol{r})\rangle &= \frac{1}{2}N_{\rm r}'(N_{\rm r}'-1)G^{\rm p}(\boldsymbol{r}), \end{aligned} \tag{3}$$

where $\xi(\boldsymbol{r})$ is the correlation function normalized to the actual number density of galaxies in the survey and

$$G^{\rm p}(\boldsymbol{r}) := \frac{2}{K^2}\sum_{i<j}^{K}\Theta^{ij}(\boldsymbol{r}) \tag{4}$$

is the fraction of microcell pairs with separation $\boldsymbol{r}$. Here $\Theta^{ij}(\boldsymbol{r}) := 1$ if $\boldsymbol{x}_i - \boldsymbol{x}_j$ falls in the $\boldsymbol{r}$-bin, otherwise it is equal to zero.

The expectation value of the LS estimator (1) is

$$\begin{aligned} \langle\hat{\xi}_{\rm LS}\rangle &= (1+\xi)\left\langle\frac{1+\alpha}{1+\gamma}\right\rangle - 2\left\langle\frac{1+\beta}{1+\gamma}\right\rangle + 1\\ &\approx \xi + (\xi-1)\langle\gamma^2\rangle + 2\langle\beta\gamma\rangle. \end{aligned} \tag{5}$$

[1] I. Zehavi, priv. comm.

A finite $R$ catalog thus introduces a (small) bias. (In LS93, $\gamma = 0$, so the estimator is unbiased in this limit). This expression is calculated to second order in $\alpha$, $\beta$, and $\gamma$ (we denote equality to second order by "≈"). Calculation to higher order is beyond the scope of this work. Since data and random catalogs are independent, $\langle\alpha\gamma\rangle = 0$.

We introduce shorthand notations $\langle\alpha_1\alpha_2\rangle$ for $\langle\alpha(\boldsymbol{r}_1)\alpha(\boldsymbol{r}_2)\rangle$, $\langle DD_1 DD_2\rangle$ for $\langle DD(\boldsymbol{r}_1)DD(\boldsymbol{r}_2)\rangle$, and similarly for other terms.

For the covariance we get

$$\begin{aligned}\mathrm{Cov}\left[\hat{\xi}_{\rm LS}(\boldsymbol{r}_1),\hat{\xi}_{\rm LS}(\boldsymbol{r}_2)\right] &\equiv \left\langle\hat{\xi}_{\rm LS}(\boldsymbol{r}_1)\hat{\xi}_{\rm LS}(\boldsymbol{r}_2)\right\rangle - \left\langle\hat{\xi}_{\rm LS}(\boldsymbol{r}_1)\right\rangle\left\langle\hat{\xi}_{\rm LS}(\boldsymbol{r}_2)\right\rangle \\ &\approx (1+\xi_1)(1+\xi_2)\langle\alpha_1\alpha_2\rangle + 4\langle\beta_1\beta_2\rangle \\ &\quad + (1-\xi_1)(1-\xi_2)\langle\gamma_1\gamma_2\rangle \\ &\quad - 2(1+\xi_1)\langle\alpha_1\beta_2\rangle - 2(1+\xi_2)\langle\beta_1\alpha_2\rangle \\ &\quad - 2(1-\xi_1)\langle\gamma_1\beta_2\rangle - 2(1-\xi_2)\langle\beta_1\gamma_2\rangle.\end{aligned} \tag{6}$$

Terms with $\gamma$ represent additional variance due to finite $N_{\rm r}'$, and are new compared to those of LS93. Also, $\langle\beta_1\beta_2\rangle$ collects an additional contribution, which we denote by $\Delta\langle\beta_1\beta_2\rangle$, from variations in the random field (see Sect. 2.3). The cross terms $\alpha_1\beta_2$ and $\alpha_2\beta_1$, instead depend linearly on the random field, and average to the $N_{\rm r}' \to \infty$ result. The additional contribution due to finite $N_{\rm r}'$ is thus

$$\begin{aligned}\Delta\mathrm{Cov}\left[\hat{\xi}_{\rm LS}(\boldsymbol{r}_1),\hat{\xi}_{\rm LS}(\boldsymbol{r}_2)\right] &\approx 4\Delta\langle\beta_1\beta_2\rangle + (1-\xi_1)(1-\xi_2)\langle\gamma_1\gamma_2\rangle \\ &\quad - 2(1-\xi_1)\langle\gamma_1\beta_2\rangle - 2(1-\xi_2)\langle\beta_1\gamma_2\rangle.\end{aligned} \tag{7}$$

From (2),

$$\langle DD_1\, DD_2\rangle = \langle DD_1\rangle\langle DD_2\rangle\left(1+\langle\alpha_1\alpha_2\rangle\right), \tag{8}$$

and so on, so that the covariances of the deviations are obtained from

$$\begin{aligned}\langle\alpha_1\alpha_2\rangle &= \frac{\langle DD_1\, DD_2\rangle - \langle DD_1\rangle\langle DD_2\rangle}{\langle DD_1\rangle\langle DD_2\rangle}, \\ \langle\beta_1\beta_2\rangle &= \frac{\langle DR_1\, DR_2\rangle - \langle DR_1\rangle\langle DR_2\rangle}{\langle DR_1\rangle\langle DR_2\rangle}, \\ \langle\gamma_1\gamma_2\rangle &= \frac{\langle RR_1\, RR_2\rangle - \langle RR_1\rangle\langle RR_2\rangle}{\langle RR_1\rangle\langle RR_2\rangle}, \\ \langle\alpha_1\beta_2\rangle &= \frac{\langle DD_1\, DR_2\rangle - \langle DD_1\rangle\langle DR_2\rangle}{\langle DD_1\rangle\langle DR_2\rangle}, \\ \langle\beta_1\gamma_2\rangle &= \frac{\langle DR_1\, RR_2\rangle - \langle DR_1\rangle\langle RR_2\rangle}{\langle DR_1\rangle\langle RR_2\rangle}, \\ \langle\alpha_1\gamma_2\rangle &= \frac{\langle DD_1\, RR_2\rangle - \langle DD_1\rangle\langle RR_1\rangle}{\langle DD_1\rangle\langle RR_2\rangle} = 0.\end{aligned} \tag{9}$$

### 2.2. Quadruplets, triplets, and approximations

We use

$$G_{12}^{\rm t} := G^{\rm t}(\boldsymbol{r}_1,\boldsymbol{r}_2) := \frac{1}{K^3}\sum_{ijk}^{*}\Theta_1^{ik}\Theta_2^{jk} \tag{10}$$

to denote the fraction of ordered microcell triplets, where $\boldsymbol{x}_i - \boldsymbol{x}_k \in \boldsymbol{r}_1$ and $\boldsymbol{x}_j - \boldsymbol{x}_k \in \boldsymbol{r}_2$. The notation $\sum^*$ means that only terms where all indices (microcells) are different are included. Here $G_{12}^{\rm t}$ is of the same magnitude as $G_1^{\rm p}\,G_2^{\rm p}$ but is larger.

Appendix A gives examples of how the $\langle DD_1\, DD_2\rangle$ and so on in (9) are calculated. These covariances involve expectation values $\langle n_i n_j n_l n_k\rangle$, where $n_i$ is the number of objects (0 or 1) in microcell $i$ and so on, and only cases where the four microcells are separated pairwise by $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are included. If all four microcells $i$, $j$, $k$, and $l$ are different, we call this case a quadruplet; it consists of two pairs with separations $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. If two of the indices, that is, microcells, are equal, we have a triplet with a center cell (the equal indices) and two end cells separated from the center by $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$.

We make the following three approximations:

1. For microcell quadruplets, the correlations between unconnected cells are approximated by zero on average.
2. Three-point correlations vanish.
3. The part of four-point correlations that does not arise from the two-point correlations vanishes.

With approximations (2) and (3), we have for the expectation value of a galaxy triplet

$$\langle n_i n_j n_k\rangle \propto 1 + \xi_{ij} + \xi_{jk} + \xi_{ik}, \tag{11}$$

where $\xi_{ij} := \xi(\boldsymbol{x}_j - \boldsymbol{x}_i)$, and for a quadruplet

$$\langle n_i n_j n_k n_l\rangle \propto 1+\xi_{ij}+\xi_{jk}+\xi_{ik}+\xi_{il}+\xi_{jl}+\xi_{kl}+\xi_{ij}\xi_{kl}+\xi_{ik}\xi_{jl}+\xi_{il}\xi_{jk}. \tag{12}$$

We use "≃" to denote results based on these three approximations. Approximation (1) is good as long as the survey size is large compared to $r_{\rm max}$. It allows us to drop terms other than $1 + \xi_{ij} + \xi_{kl} + \xi_{ij}\xi_{kl}$ in (12). Approximations (2) and (3) hold for Gaussian density fluctuations, but in the realistic cosmological situation they are not good: the presence of the higher-order correlations makes the estimation of the covariance of $\xi(\boldsymbol{r})$ estimators a difficult problem. However, this difficulty applies only to the contribution of the data to the covariance, that is, to the part that does not depend on the size and treatment of the random catalog. The key point in this work is that while our theoretical result for the total covariance does not hold in a realistic situation (it is an underestimate), our results for the difference in estimator covariance due to different treatments of the random catalog hold well.

In addition to working in the limit $N_{\rm r}' \to \infty$ ($\gamma = 0$), LS93 considered only 1D bins and the case where $r_1 = r_2 \equiv r$ (i.e., variances, not covariances) and made also a fourth approximation: for triplets (which in this case have legs of equal length) they approximated the correlation between the end cells (whose separation in this case varies between 0 and $2r$) by $\xi(r)$. We use $\xi_{12}$ to denote the mean value of the correlation between triplet end cells (separated from the triplet center by $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$). (For our plots in Sect. 4 we make a similar approximation of $\xi_{12}$ as Landy & Szalay 1993, see Sect. 4.2). Also, LS93 only calculated to first order in $\xi$, whereas we do not make this approximation. Bernstein (1994) also considered covariances, and included the effect of three-point and four-point correlations, but worked in the limit $N_{\rm r}' \to \infty$ ($\gamma = 0$).

### 2.3. Poisson, edge, and $q$ terms

After calculating all the $\langle DD_1\, DD_2\rangle$ and so on (see Appendix A), (9) becomes

$$\begin{aligned}&(1+\xi_1)(1+\xi_2)\langle\alpha_1\alpha_2\rangle \\ &\quad\simeq \frac{4}{N_{\rm d}}(1+\xi_1)(1+\xi_2)\left[\frac{G_{12}^{\rm t}}{G_1^{\rm p}\,G_2^{\rm p}} - 1\right] \\ &\qquad + \frac{2(1+\xi_1)}{N_{\rm d}(N_{\rm d}-1)}\left[\frac{\delta_{12}}{G_1^{\rm p}} - 2(1+\xi_2)\frac{G_{12}^{\rm t}}{G_1^{\rm p}\,G_2^{\rm p}} + (1+\xi_2)\right] \\ &\qquad + \frac{4(N_{\rm d}-2)}{N_{\rm d}(N_{\rm d}-1)}(\xi_{12}-\xi_1\xi_2)\frac{G_{12}^{\rm t}}{G_1^{\rm p}\,G_2^{\rm p}},\end{aligned}$$

$$\begin{aligned}
\langle\beta_1\beta_2\rangle &\simeq \frac{1}{N_\mathrm{d}N_\mathrm{r}}\left\{N'_\mathrm{r}\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right]+N_\mathrm{d}\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right]+1\right.\\
&\quad\left.-\frac{2G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}+\frac{\delta_{12}}{G^\mathrm{p}_1}\right\}+\frac{N_\mathrm{d}-1}{N_\mathrm{d}N_\mathrm{r}}\xi_{12}\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2},\\
\langle\gamma_1\gamma_2\rangle &= \frac{2}{N'_\mathrm{r}(N'_\mathrm{r}-1)}\left\{2(N'_\mathrm{r}-2)\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right]+\frac{\delta_{12}}{G^\mathrm{p}_1}-1\right\},\\
\langle\alpha_1\beta_2\rangle &\simeq \frac{2}{N_\mathrm{d}}\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right],\\
\langle\beta_1\gamma_2\rangle &= \frac{2}{N'_\mathrm{r}}\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right],\\
\langle\alpha_1\gamma_2\rangle &= 0,
\end{aligned}\tag{13}$$

for the standard LS estimator.

Following the definition of $t$ and $p$ in LS93, we define

$$\begin{aligned}
t_{12} &:= \frac{1}{N_\mathrm{d}}\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right],\\
t^\mathrm{r}_{12} &:= \frac{1}{N'_\mathrm{r}}\left[\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}-1\right] = \frac{N_\mathrm{d}}{N'_\mathrm{r}}\,t_{12},\\
p_{12} &:= \frac{2}{N_\mathrm{d}(N_\mathrm{d}-1)}\left[\frac{\delta_{12}}{(1+\xi_1)G^\mathrm{p}_1}-2\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}+1\right],\\
p^\mathrm{c}_{12} &:= \frac{1}{N_\mathrm{d}N_\mathrm{r}}\left[\frac{\delta_{12}}{G^\mathrm{p}_1}-2\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}+1\right],\\
p^\mathrm{r}_{12} &:= \frac{2}{N'_\mathrm{r}(N'_\mathrm{r}-1)}\left[\frac{\delta_{12}}{G^\mathrm{p}_1}-2\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2}+1\right],\\
q_{12} &:= \frac{1}{N_\mathrm{d}}\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2} = t_{12}+\frac{1}{N_\mathrm{d}},\\
q^\mathrm{r}_{12} &:= \frac{1}{N'_\mathrm{r}}\frac{G^\mathrm{t}_{12}}{G^\mathrm{p}_1G^\mathrm{p}_2} = t^\mathrm{r}_{12}+\frac{1}{N'_\mathrm{r}}.
\end{aligned}\tag{14}$$

For their diagonals ($\boldsymbol{r}_1=\boldsymbol{r}_2$), we write $t$, $t_\mathrm{r}$, $p$, $p_\mathrm{c}$, $p_\mathrm{r}$, $q$, and $q_\mathrm{r}$. Thus, $t\equiv t_{11}\equiv t_{22}$, $t_\mathrm{r}\equiv t^\mathrm{r}_{11}\equiv t^\mathrm{r}_{22}$ and so on. (We use superscripts for the matrices, e.g., $t^\mathrm{r}(\boldsymbol{r}_1,\boldsymbol{r}_2)$, and subscripts for their diagonals, e.g., $t_\mathrm{r}(\boldsymbol{r})$).

Using these definitions, (13) becomes

$$\begin{aligned}
(1+\xi_1)(1+\xi_2)\langle\alpha_1\alpha_2\rangle &\simeq (1+\xi_1)(1+\xi_2)(4t_{12}+p_{12})\\
&\quad+4\frac{N_\mathrm{d}-2}{(N_\mathrm{d}-1)}(\xi_{12}-\xi_1\xi_2)q_{12},\\
\langle\beta_1\beta_2\rangle &\simeq t_{12}+t^\mathrm{r}_{12}+p^\mathrm{c}_{12}+\frac{N_\mathrm{d}-1}{N_\mathrm{d}}\xi_{12}q^\mathrm{r}_{12},\\
\langle\gamma_1\gamma_2\rangle &= 4\,t^\mathrm{r}_{12}+p^\mathrm{r}_{12},\\
\langle\alpha_1\beta_2\rangle &\simeq 2t_{12},\ \langle\beta_1\gamma_2\rangle = 2\,t^\mathrm{r}_{12},\quad\text{and}\quad\langle\alpha_1\gamma_2\rangle = 0.
\end{aligned}\tag{15}$$

The new part in $\langle\beta_1\beta_2\rangle$ due to finite size of the random catalog is

$$\Delta\langle\beta_1\beta_2\rangle \simeq t^\mathrm{r}_{12}+p^\mathrm{c}_{12}+\frac{N_\mathrm{d}-1}{N_\mathrm{d}}\xi_{12}q^\mathrm{r}_{12}.\tag{16}$$

Thus only $\langle\alpha_1\alpha_2\rangle$ and $\langle\beta_1\beta_2\rangle$ are affected by $\xi(\boldsymbol{r})$ (in our approximation its effect cancels in $\langle\alpha_1\beta_2\rangle$). The results for $\langle\gamma_1\gamma_2\rangle$, $\langle\beta_1\gamma_2\rangle$, and $\langle\alpha_1\gamma_2\rangle$ are exact. The result for $\langle\alpha_1\alpha_2\rangle$ involves all three approximations mentioned above, $\langle\alpha_1\beta_2\rangle$ involves approximations (1) and (2), and $\langle\beta_1\beta_2\rangle$ involves approximation (1).

We refer to $p$, $p^\mathrm{c}$, and $p^\mathrm{r}$ as "Poisson" terms and $t$ and $t^\mathrm{r}$ as "edge" terms (the difference between $G^\mathrm{t}_{12}$ and $G^\mathrm{p}_1G^\mathrm{p}_2$ is due to edge effects). While the Poisson terms are strongly diagonal dominated, the edge terms are not. Since $N_\mathrm{d}t_{12}=N'_\mathrm{r}t^\mathrm{r}_{12}\ll 1$, the $q$ terms are much larger than the edge terms, but they get multiplied by $\xi_{12}-\xi_1\xi_2$ or $\xi_{12}$. In the limit $N'_\mathrm{r}\to\infty$: $\langle\beta_1\gamma_2\rangle\to 0$, $\langle\gamma_1\gamma_2\rangle\to 0$, $\langle\beta_1\beta_2\rangle\to t_{12}$; $\langle\alpha_1\alpha_2\rangle$, and also $\langle\alpha_1\beta_2\rangle$ are unaffected.

We see that *DD–DR* and *DR–RR* correlations arise from edge effects. If we increase the density of data or random objects, the Poisson terms decrease as $N^{-2}$ but the edge terms decrease only as $N^{-1}$ so the edge effects are more important for a higher density of objects.

Doubling the bin size (combining neighboring bins) doubles $G^\mathrm{p}(\boldsymbol{r})$ but makes $G^\mathrm{t}(\boldsymbol{r}_1,\boldsymbol{r}_2)$ four times as large, since triplets where one leg was in one of the original smaller bins and the other leg was in the other bin are now also included. Thus, the ratio $G^\mathrm{t}_{12}/(G^\mathrm{p}_1G^\mathrm{p}_2)$ and $t$ are not affected, but the dominant term in $p$, $1/(1+\xi)G^\mathrm{p}$ is halved. Edge effects are thus more important for larger bins.

### 2.4. Results for the standard Landy–Szalay estimator

Inserting the results for $\langle\alpha_1\alpha_2\rangle$ and so on into Eqs. (5) and (6), we get that the expectation value of the standard LS estimator (1) is

$$\langle\hat{\xi}_\mathrm{LS}\rangle = \xi+(\xi-1)\,(4\,t_\mathrm{r}+p_\mathrm{r})+4\,t_\mathrm{r}.\tag{17}$$

This holds also for large $\xi$ and in the presence of three-point and four-point correlations. A finite $R$ catalog thus introduces a bias $(\xi-1)\,(4\,t_\mathrm{r}+p_\mathrm{r})+4\,t_\mathrm{r}=-p_\mathrm{r}+(4\,t_\mathrm{r}+p_\mathrm{r})\xi$; the edge ($t_\mathrm{r}$) part of the bias cancels in the $\xi\to 0$ limit.

For the covariance we get

$$\begin{aligned}
\mathrm{Cov}\left[\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_1),\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_2)\right] &\equiv \left\langle\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_1)\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_2)\right\rangle-\left\langle\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_1)\right\rangle\left\langle\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_2)\right\rangle\\
&\simeq (1+\xi_1)(1+\xi_2)p_{12}+4p^\mathrm{c}_{12}\\
&\quad+(1-\xi_1)(1-\xi_2)\,p^\mathrm{r}_{12}+4\xi_1\xi_2(t_{12}+t^\mathrm{r}_{12})\\
&\quad+4\frac{N_\mathrm{d}-2}{N_\mathrm{d}-1}(\xi_{12}-\xi_1\xi_2)q_{12}+4\frac{N_\mathrm{d}-1}{N_\mathrm{d}}\xi_{12}q^\mathrm{r}_{12}.
\end{aligned}\tag{18}$$

Because of the approximations made, this result for the covariance does not apply to the realistic cosmological case; not even for large separations $r$, where $\xi$ is small, since large correlations at small $r$ increase the covariance also at large $r$. However, this concerns only $\langle\alpha_1\alpha_2\rangle$ and $\langle\alpha_1\beta_2\rangle$. Our focus here is on the additional covariance due to the size and handling of the random catalog, which for standard LS is

$$\begin{aligned}
\Delta\mathrm{Cov}\left[\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_1),\hat{\xi}_\mathrm{LS}(\boldsymbol{r}_2)\right] &\approx 4\Delta\langle\beta_1\beta_2\rangle+(1-\xi_1)(1-\xi_2)\langle\gamma_1\gamma_2\rangle\\
&\quad-2(1-\xi_1)\langle\gamma_1\beta_2\rangle-2(1-\xi_2)\langle\beta_1\gamma_2\rangle\\
&\simeq 4p^\mathrm{c}_{12}+(1-\xi_1)(1-\xi_2)\,p^\mathrm{r}_{12}+4\xi_1\xi_2\,t^\mathrm{r}_{12}\\
&\quad+4\frac{N_\mathrm{d}-1}{N_\mathrm{d}}\xi_{12}q^\mathrm{r}_{12}.
\end{aligned}\tag{19}$$

To zeroth order in $\xi$ the covariance is given by the Poisson terms and the edge terms cancel to first order in $\xi$. This is the property for which the standard LS estimator was designed. To first order in $\xi$, the $q$ terms contribute. This $q$ contribution involves the triplet correlation $\xi_{12}$, which, depending on the form of $\xi(\boldsymbol{r})$, may be larger than $\xi_1$ or $\xi_2$.

If we try to save cost by using a diluted random catalog with $N'_{\rm r} \ll N_{\rm r}$ for $RR$ pairs, $\langle\gamma_1\gamma_2\rangle$ is replaced by $\langle\gamma'_1\gamma'_2\rangle = 4\,t^{\rm r'}_{12} + p^{\rm r''}_{12}$ with $N'_{\rm r}$ in place of $N_{\rm r}$, but $\langle\beta_1\gamma'_2\rangle = \langle\beta_1\gamma_2\rangle$ and $\langle\beta_1\beta_2\rangle$ are unaffected, so that the edge terms involving randoms no longer cancel. In Sect. 4 we see that this is a large effect. Therefore, one should not use dilution.

## 3. Split random catalog

### 3.1. Bias and covariance for the split method

In the split method one has $M_{\rm s}$ independent smaller $R^\mu$ catalogs of size $N'_{\rm r}$ instead of one large random catalog $R$. Their union, $R$, has a size of $N'_{\rm r} = M_{\rm s}N'_{\rm r}$. The pair counts $DR(\boldsymbol{r})$ and $RR'(\boldsymbol{r})$ are calculated as

$$DR(\boldsymbol{r}) := \sum_{\mu=1}^{M_{\rm s}} DR^\mu(\boldsymbol{r}) \quad \text{and} \quad RR'(\boldsymbol{r}) := \sum_{\mu=1}^{M_{\rm s}} R^\mu R^\mu(\boldsymbol{r}), \tag{20}$$

that is, pairs across different $R^\mu$ catalogs are not included in $RR'$. The total number of pairs in $RR'$ is $\frac{1}{2}M_{\rm s}N'_{\rm r}(N'_{\rm r}-1) = \frac{1}{2}N'_{\rm r}(N'_{\rm r}-1)$. Here, $DR$ is equal to its value in standard LS.

The split Landy–Szalay estimator is

$$\hat{\xi}_{\rm split}(\boldsymbol{r}) := \frac{N'_{\rm r}(N'_{\rm r}-1)}{N_{\rm d}(N_{\rm d}-1)}\frac{DD(\boldsymbol{r})}{RR'(\boldsymbol{r})} - \frac{N'_{\rm r}-1}{N_{\rm d}}\frac{DR(\boldsymbol{r})}{RR'(\boldsymbol{r})} + 1\cdot \tag{21}$$

Compared to standard LS, $\langle\alpha_1\alpha_2\rangle$, $\langle\beta_1\beta_2\rangle$, and $\langle\alpha_1\beta_2\rangle$ are unaffected. We construct $\langle RR'\rangle$, $\langle RR'\cdot RR'\rangle$, and $\langle RR'\cdot DR\rangle$ from the standard LS results, bearing in mind that the random catalog is a union of independent catalogs, arriving at

$$\begin{aligned}\langle\beta_1\gamma'_2\rangle &= 2\,t^{\rm r}_{12},\\ \langle\gamma'_1\gamma'_2\rangle &= 4\,t^{\rm r}_{12} + p^{\rm r'}_{12},\end{aligned} \tag{22}$$

where

$$p^{\rm r'}_{12} := \frac{N_{\rm d}(N_{\rm d}-1)}{N'_{\rm r}(N'_{\rm r}-1)}\,p_{12} \equiv \frac{N'_{\rm r}-1}{N'_{\rm r}-1}\,p^{\rm r}_{12}. \tag{23}$$

The first is the same as in standard LS and dilution, but the second differs both from standard LS and from dilution, since it involves both $N'_{\rm r}$ and $N'_{\rm r}$.

For the expectation value we get

$$\langle\hat{\xi}_{\rm split}\rangle = \xi + (\xi-1)\,(4\,t_{\rm r} + p'_{\rm r}) + 4\,t_{\rm r}, \tag{24}$$

so that the bias is $(\xi-1)(4\,t_{\rm r}+p'_{\rm r}) + 4\,t_{\rm r} = -p'_{\rm r} + (4\,t_{\rm r}+p'_{\rm r})\xi$. In the limit $\xi \to 0$ the edge part cancels, leaving only the Poisson term.

The covariance is

$$\begin{aligned}{\rm Cov}\left[\hat{\xi}_{\rm split}(\boldsymbol{r}_1), \hat{\xi}_{\rm split}(\boldsymbol{r}_2)\right] \simeq{}& (1+\xi_1)(1+\xi_2)p_{12} + 4\,p^{\rm c}_{12}\\ &+ (1-\xi_1)(1-\xi_2)p^{\rm r'}_{12} + 4\xi_1\xi_2(t_{12} + t^{\rm r}_{12}).\end{aligned} \tag{25}$$

The *change* in the covariance compared to the standard LS method is

$${\rm Cov}\left[\hat{\xi}^{\rm split}_1, \hat{\xi}^{\rm split}_2\right] - {\rm Cov}\left[\hat{\xi}^{\rm LS}_1, \hat{\xi}^{\rm LS}_2\right] = (1-\xi_1)(1-\xi_2)(p^{\rm r'}_{12} - p^{\rm r}_{12}), \tag{26}$$

which again applies in the realistic cosmological situation. Our main result is that in the split method the edge effects cancel and the bias and covariance are the same as for standard LS, except that the Poisson term $p^{\rm r}$ from $RR$ is replaced with the larger $p^{\rm r'}$.

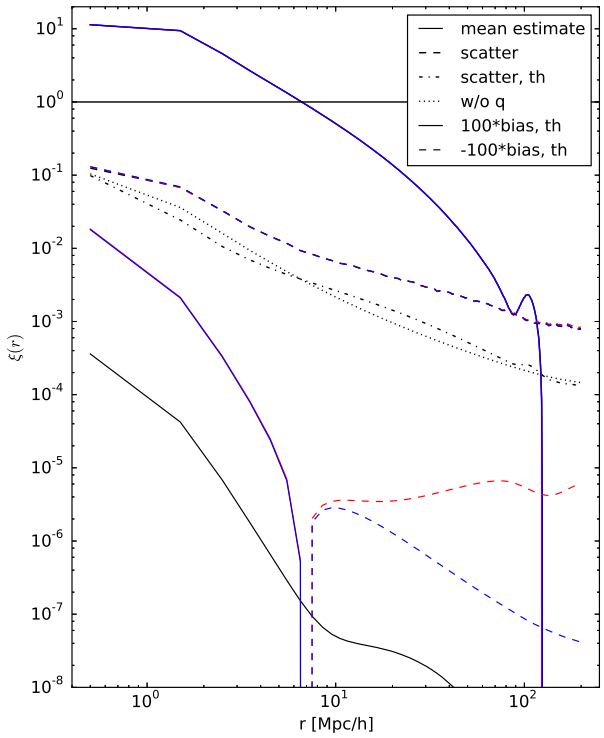


**Fig. 1.** Mean $\xi(r)$ estimate and the scatter and theoretical bias of the estimates for different estimators. The dash-dotted line, our theoretical result for the scatter of the LS method, underestimates the scatter, since higher-order correlations in the $D$ catalog are ignored. The dotted line is without the contribution of the $q$ terms, and is dominated by the Poisson ($p$) terms. The bias is multiplied by 100 so the curves can be displayed in a more compact plot. For the measured mean and scatter, and the theoretical bias we plot standard LS in black, dilution with $d = 0.14$ in red, and split with $M_{\rm s} = 50$ in blue. For the mean and scatter the difference between the methods is not visible in this plot. The differences in the mean estimate are shown in Fig. 5. The differences in scatter (or its square, the variance) are shown in Fig 3. For the theoretical bias the difference between split and dilution is not visible at small $r$ ($\xi(r) > 1$), where the bias is positive.

### 3.2. Optimizing computational cost and variance of the split method

The bias is small compared to variance in our application (see Fig. 1 for the theoretical result and Fig. 5 for an attempted bias measurement), and therefore we focus on variance as the figure of merit. The computational cost should be roughly proportional to

$$\frac{1}{2}N_{\rm d}^2\left(1 + 2M_{\rm r} + \frac{M_{\rm r}^2}{M_{\rm s}}\right) =: \frac{1}{2}N_{\rm d}^2 c, \tag{27}$$

and the additional variance due to finite $R$ catalog in the $\xi \to 0$ limit becomes

$$\Delta{\rm var} \approx \left(\frac{2}{M_{\rm r}} + \frac{M_{\rm s}}{M_{\rm r}^2}\right)p =: vp. \tag{28}$$

Here, $N_{\rm d}$ and $p$ are fixed by the survey and the requested $\boldsymbol{r}$ binning, but we can vary $M_{\rm r}$ and $M_{\rm s}$ in the search for the optimal computational method. In the above we defined the "cost" and "variance" factors $c$ and $v$.

We may ask two questions:

1. For a fixed level of variance $v$, which combination of $M_{\rm r}$ and $M_{\rm s}$ minimizes computational cost $c$?

2. For a fixed computational cost $c$, which combination of $M_r$ and $M_s$ minimizes the variance $v$?

The answer to both questions is (Slepian & Eisenstein 2015)

$$M_s = M_r \quad \Rightarrow \quad c = 1 + 3M_r \quad \text{and} \quad v = \frac{3}{M_r}\cdot \tag{29}$$

Thus, the optimal version of the split method is the natural one where $N'_r = N_d$. In this case the additional variance in the $\xi \to 0$ limit becomes

$$\Delta\text{var} \approx \left(2\frac{N_d}{N'_r} + \frac{N_d}{N'_r}\right)p, \tag{30}$$

and the computational cost factor $N_d^2 + 2N_dN_r + N'_rN'_r$ becomes

$$\left(1 + 2\frac{N'_r}{N_d} + \frac{N'_r}{N_d}\right)N_d^2, \tag{31}$$

meaning that *DR* pairs contribute twice as much as *RR* pairs to the variance and also twice as much computational cost is invested in them. The memory requirement for the random catalog is then the same as for the data catalog. The cost saving estimate above is optimistic, since the computation involves some overhead not proportional to the number of pairs.

For small scales, where $\xi \gg 1$, the situation is different. The greater density of *DD* pairs due to the correlation requires a greater density of the *R* catalog so that the additional variance from it is not greater. From Eq. (19) we see that the balance of the *DR* and the *RR* contributions is different for large $\xi$ (the $p_c$ term vs. the other terms). We may consider recomputing $\hat{\xi}$ for the small scales using a smaller $r_{max}$ and a larger *R* catalog. Considering just the Poisson terms ($p_c$ and $p_r$ or $p'_r$) with a "representative" $\xi$ value, (27) and (28) become $c = 1 + \xi + 2\,M_r + M_r^2/M_s$ and $v = 2/M_r + (1-\xi)^2 M_s/M_r^2$ which modifies the above result (Eq. (29)) for the optimal choice of $M_s$ and $M_r$ to

$$M_s = \frac{M_r}{|\xi - 1|}, \quad \text{that is,} \quad N'_r = |\xi - 1|N_d. \tag{32}$$

This result is only indicative, since it assumes a constant $\xi$ for $r < r_{max}$. In particular, it does not apply for $\xi \approx 1$, because then the approximation of ignoring the $q^r$ and $t^r$ terms in (19) is not good.

## 4. Tests on mock catalogs

### 4.1. Minerva simulations and methodology

The Minerva mocks are a set of 300 cosmological mocks produced with N-body simulations (Grieb et al. 2016; Lippich et al. 2019), stored at five output redshifts $z \in \{2.0, 1.0, 0.57, 0.3, 0\}$. The cosmology is flat ΛCDM with $\Omega_m = 0.285$, and we use the $z = 1$ outputs. The mocks have $N_d \approx 4 \times 10^6$ objects ("halos" found by a friends-of-friend algorithm) in a box of $1500\,h^{-1}$ Mpc cubed.

To model the survey geometry of a redshift bin with $\Delta z \approx 0.1$ at $z \sim 1$, we placed the observer at comoving distance $2284.63\,h^{-1}$ Mpc from the center of the cube and selected from the cube a shell 2201.34–2367.92 $h^{-1}$ Mpc from the observer. The comoving thickness of the shell is $166.58\,h^{-1}$ Mpc. The resulting mock sub-catalogs have $N_d \approx 4.5 \times 10^5$ and are representative of the galaxy number density of the future Euclid spectroscopic galaxy catalog.

We ignore peculiar velocities, that is, we perform our analysis in real space. Therefore, we consider results for the 1D 2PCF $\xi(r)$. We estimated $\xi(r)$ up to $r_{max} = 200\,h^{-1}$ Mpc using $\Delta r = 1\,h^{-1}$ Mpc bins.

We chose standard LS with $M_r = 50$ as the reference method. In the following, LS without further qualification refers to this. The random catalog was generated separately for each shell mock to measure their contribution to the variance. For one of the random catalogs we calculated also triplets to obtain the edge effect quantity $N_d t_{12} = G^t_{12}/G^p_1 G^p_2 - 1$.

While dilution can already be discarded on theoretical grounds, we show results obtained using dilution, since these results provide the scale for edge effects demonstrating the importance of eliminating them with a careful choice of method. For the dilution and split methods we also used $M_r = 50$, and tried out dilution fractions $d := N'_r/N'_r = 0.5, 0.25, 0.14$ and split factors $M_s = 4, 16, 50$ (chosen to have similar pairwise computational costs). In addition, we considered standard LS with $M_r = 25$, which has the same number of *RR* pairs as $d = 0.5$ and $M_s = 4$, but only half the number of *DR* pairs; and standard LS with $M_r = 1$ to demonstrate the effect of a small $N'_r$.

The code used to estimate the 2PCF implements a highly optimized pair-counting method, specifically designed for the search of object pairs in a given range of separations. In particular, the code provides two alternative counting methods, the *chain-mesh* and the *kd-tree*. Both methods measure the exact number of object pairs in separation bins, without any approximation. However, since they implement different algorithms to search for pairs, they perform differently at different scales, both in terms of CPU time and memory usage. Overall, the efficiency of the two methods depends on the ratio between the scale range of the searching region and the maximum separation between the objects in the catalog.

The *kd-tree* method first constructs a space-partitioning data structure that is filled with catalog objects. The minimum and maximum separations probed by the objects are kept in the data structure and are used to prune object pairs with separations outside the range of interest. The tree pair search is performed through the dual-tree method in which cross-pairs between two dual trees are explored. This is an improvement in terms of exploration time over the single-tree method.

On the other hand, in the *chain-mesh* method the catalog is divided in cubic cells of equal size, and the indexes of the objects in each cell are stored in vectors. To avoid counting object pairs with separations outside the interest range, the counting is performed only on the cells in a selected range of distances from each object. The *chain-mesh* algorithm has been imported from the CosmoBolognaLib, a large set of *free software* C++/python libraries for cosmological calculations (Marulli et al. 2016).

For our test runs we used the *chain-mesh* method.

### 4.2. Variance and bias

In Fig. 1 we show the mean (over the 300 mock shells) estimated correlation function and the scatter (square root of the variance) of the estimates using the LS, split, and dilution methods; our theoretical approximate result for the scatter for LS; and our theoretical result for bias for the different methods.

The theoretical result for the scatter is shown with and without the $q$ terms, which include the triplet correlation $\xi_{12}$, for which we used here the approximation $\xi_{12} \approx \xi(\max(r_1, r_2))$. This behaves as expected, that is, it underestimates the variance, since we neglected the higher-order correlations in the *D* catalog. Nevertheless, it (see the dash-dotted line in Fig. 1) has similar features to the measured variance (dashed lines).

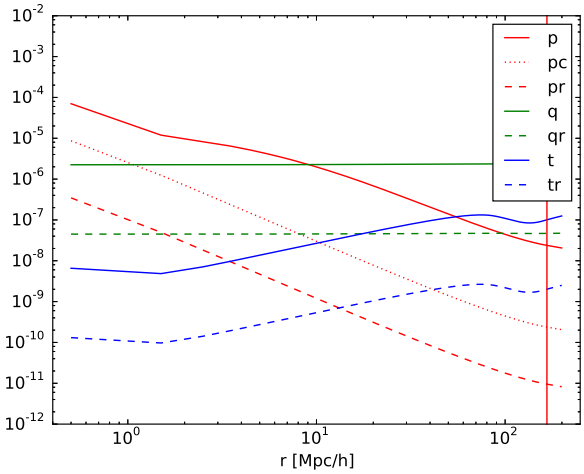


**Fig. 2.** Quantities $p$, $p_c$, $p_r$, $q$, $q_r$, $t$, and $t_r$ for the Minerva shell. The values for the first bin are noisy. The vertical red line marks $r = L$.

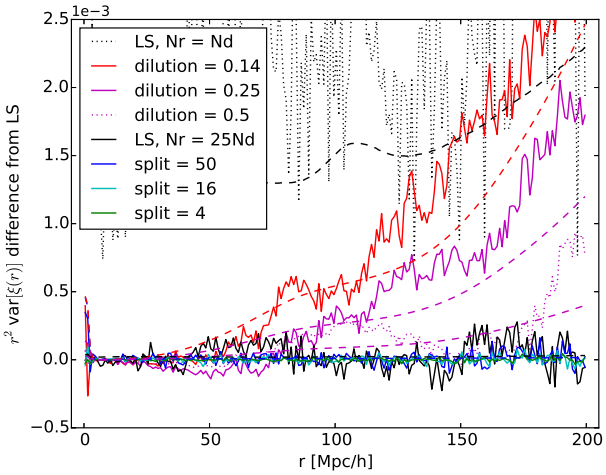


**Fig. 3.** Measured difference from LS of the variance of different estimators, multiplied by $r^2$. Dashed lines are our theoretical results.

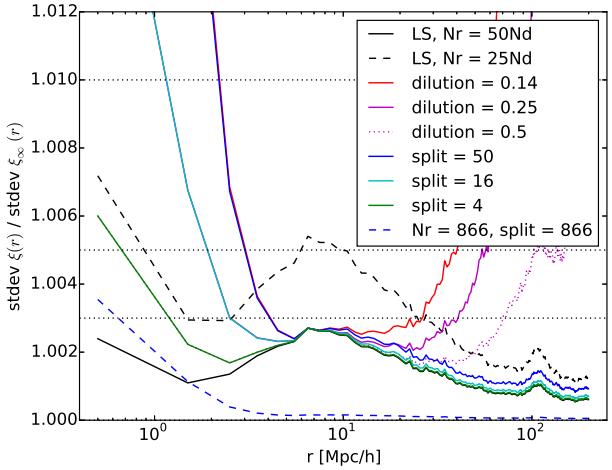


**Fig. 4.** Theoretical estimate of the scatter of the $\xi$ estimates divided by the scatter in the $N'_r \to \infty$ limit. The dotted lines correspond to 0.3%, 0.5%, and 1% increase in scatter. For $r < 10\,h^{-1}$ Mpc there is hardly any difference between split and dilution, the curves lie on top of each other; whereas for larger $r$ split is much better.

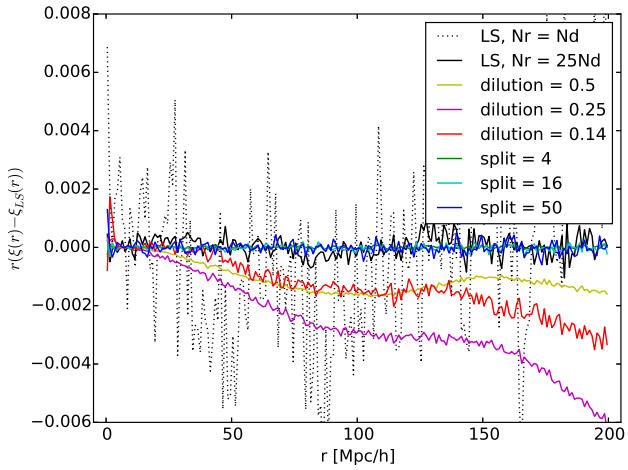


**Fig. 5.** Differences between the mean $\xi(r)$ estimate and that from the LS, multiplied by $r$ to better display all scales. This measured difference is not the true bias, which is too small to measure with 300 mocks, and is mainly due to random error of the mean. The results for dilution appear to reveal a systematic bias, but this is just due to strong error correlations between nearby bins; for different subsets of the 300 mocks the mean difference is completely different.

In Fig. 2 we plot the diagonals of the $p$, $t$, and $q$ quantities. This shows how their relative importance changes with separation scale. It also confirms that our initial assumption on small relative fluctuations is valid in this simulation case.

Consider now the variance differences (from standard LS with $M_r = 50$), for which our theoretical results should be accurate. Figure 3 compares the measured variance difference to the theoretical result. For the diluted estimators and LS with $M_r = 1$ the measured result agrees with theory, although clearly the measurement with just 300 mocks is rather noisy. For the split estimators and LS with $M_r = 25$ the difference is too small to be appreciated with 300 mocks, but at least the measurement does not disagree with the theoretical result.

In Fig. 4 we show the relative theoretical increase in scatter compared to the best possible case, which is LS in the limit $M_r \to \infty$. Since we do not have a valid theoretical result for the total scatter, we estimate it by subtracting the theoretical difference from LS with $M_r = 50$ from the measured variance of the latter.

At scales $r \lesssim 10\,h^{-1}$ Mpc the theoretical prediction is about the same for dilution and split and neither method looks promising for $r \ll 10\,h^{-1}$ Mpc where $\xi \gg 1$. This suggests that for optimizing cost and accuracy, a different method should be used for smaller scales than that used for large scales. The number of *RR* pairs with small separations is much less. Therefore, for the small-scale computation there is no need to restrict the computation to a subset of *RR* pairs, or alternatively, one can afford to increase $M_r$. For the small scales, we may consider the split method with increased $M_r$ as an alternative to standard LS. We have the same number of pairs to compute as in the reference LS case, if we use $M_r = 866$ and $M_s = 866$. We added this case to Fig. 4. It seems to perform better than LS at intermediate scales, but for the smallest scales LS has the smaller variance. This is in line with our conclusion in Sect. 3.2, which is that when $\xi \gg 1$, it is not optimal to split the *R* catalog into small subsets.

We also compared the differences in the mean estimate from the different estimators to our theoretical results on the bias differences (see Fig. 5), but the theoretical bias differences are much smaller than the expected error of the mean from

**Table 1.** Mean computation time over the 300 runs and the mean variance over four different ranges of $r$ bins (given in units of $h^{-1}$ Mpc) for each method.

| Method | Time | Mean variance | | | |
|---|---|---|---|---|---|
| | [s] | 0–2 [$\times 10^{-2}$] | 5–15 [$\times 10^{-5}$] | 80–120 [$\times 10^{-6}$] | 150–200 [$\times 10^{-7}$] |
| $M_r = 50$ | 7889 | 1.01 | 4.42 | 1.23 | 7.01 |
| $M_r = 25$ | 2306 | 1.01 | 4.39 | 1.23 | 7.04 |
| $M_r = 1$ | 16.5 | 5.96 | 5.53 | 1.46 | 8.05 |
| $d = 0.5$ | 2239 | 1.01 | 4.41 | 1.25 | 7.11 |
| $d = 0.25$ | 812 | 1.02 | 4.41 | 1.25 | 7.42 |
| $d = 0.14$ | 487 | 1.08 | 4.41 | 1.28 | 7.72 |
| $M_s = 4$ | 1854 | 1.01 | 4.42 | 1.23 | 7.02 |
| $M_s = 16$ | 763 | 1.00 | 4.42 | 1.23 | 7.01 |
| $M_s = 50$ | 593 | 1.09 | 4.42 | 1.23 | 7.02 |

**Notes.** The first three are standard LS. The variance cannot be measured accurately enough from 300 realizations to correctly show all the differences between methods. Thus the table shows some apparent improvements (going from $M_r = 50$ to $M_r = 25$ and from $M_s = 4$ to $M_s = 16$), which are not to be taken as real. See Fig. 6 for the fifth vs. second column with error bars.

300 mocks; and we simply confirm that the differences we see are consistent with the error of the mean and thus consistent with the true bias being much smaller. We also performed tests with completely random ($\xi = 0$) mocks, and with a large number (10 000) of mocks confirmed the theoretical bias result for the different estimators in this case. Since the bias is too small to be interesting we do not report these results in more detail here.

However, we note that for the estimation of the 2D 2PCF and its multipoles, the 2D bins will contain a smaller number of objects than the 1D bins of these test runs and therefore the bias is larger. Using the theoretical results (17) or (24) the bias can be removed afterwards with accuracy depending on how well we know the true $\xi$.

### 4.3. Computation time and variance

The test runs were made using a single full 24-core node for each run. Table 1 shows the mean computation time and mean estimator variance for different $r$ ranges for the different cases we tested. Of these $r$ ranges, the $r = 80{-}120\,h^{-1}$ Mpc is perhaps the most interesting, since it contains the important BAO scale. Therefore, we plot the mean variance at this range versus mean computation time in Fig. 6, together with our theoretical predictions. The theoretical estimate for the computation time for other dilution fractions and split factors is

$$(1 + 24.75 d^2)\ 306\,\mathrm{s} \quad \text{and} \quad (1 + 24.75/M_s^2)\ 306\,\mathrm{s}, \tag{33}$$

assuming $M_r = 50$. For standard LS with other random catalog sizes, the computation time estimate is

$$(1 + 2\,M_r + M_r^2)\ 3.03\,\mathrm{s}. \tag{34}$$

## 5. Conclusions

The computational time of the standard Landy–Szalay estimator is dominated by the *RR* pairs. However, except at small scales where correlations are large, these make a negligible contribution to the expected error compared to the contribution from the

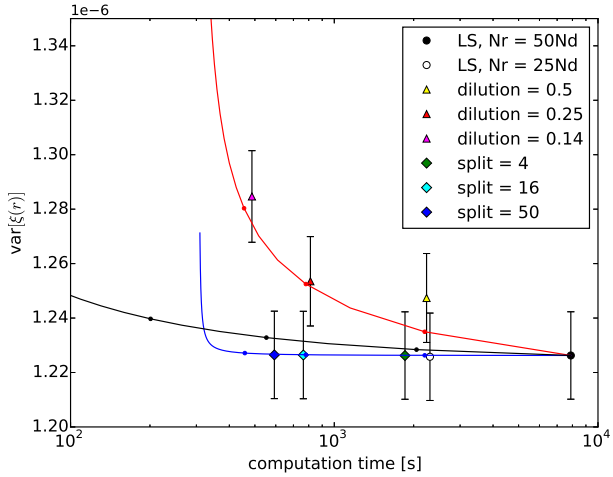


**Fig. 6.** Measured variance (mean variance over the range $r = 80{-}120\,h^{-1}$ Mpc) vs. computational cost (mean computation time) for the different methods (markers with error bars) and our theoretical prediction (solid lines). The solid lines (blue for the split method, red for dilution, and black for standard LS with $M_r \leq 50$) are our theoretical predictions for the increase in variance and computation time ratio when compared to the standard LS, $M_r = 50$, case, and the dots on the curves correspond to the measured cases (except for LS they are, from right to left, $M_r = 25$, 12.5, and (50/7); only the first of which was measured). The curve for split ends at $M_s = 2500$; the optimal case, $M_s = M_r$, is the circled dot. The error bars for the variance measurement are naive estimates that do not account for error correlations between bins. The theoretical predictions overestimate the cost savings (data points are to the right of the dots on curves; except for the smaller split factors, where the additional speed-up compared to theory is related to some other performance differences between our split and standard LS implementations). This plot would have a different appearance for other $r$ ranges.

*DD* and *DR* pairs. Therefore, a substantial saving of computation time with an insignificant loss of accuracy can be achieved by counting a smaller subset of *RR* pairs.

We considered two ways to reduce the number of *RR* pairs: dilution and split. In dilution, only a subset of the *R* catalog is used for *RR* pairs. In split, the *R* catalog is split into a number of smaller subcatalogs, and only pairs within each subcatalog are counted. We derived theoretical results for the additional estimator covariance and bias due to the finite size of the random catalog for these different variants of the LS estimator, extending in many ways the original results by Landy & Szalay (1993), who worked in the limit of an infinite random catalog. We tested our results using 300 mock data catalogs, representative of the $z = 0.95{-}1.05$ redshift range of the future Euclid survey. The split method maintains the property the Landy–Szalay estimator was designed for, namely cancelation of edge effects in bias and variance (for $\xi = 0$), whereas dilution loses this cancellation and therefore should not be used.

For small scales, where correlations are large, one should not reduce *RR* counts as much. The natural dividing line is the scale $r$ where $\xi(r) = 1$. Interestingly, the difference in bias and covariance between the different estimators (split, dilution, and LS) vanishes when $\xi = 1$. We recommend the natural version of the split method, $M_s = M_r$, for large scales where $|\xi| < 1$. This leads to a saving in computation time by more than a factor of ten (assuming $M_r = 50$) with a negligible effect on variance and bias. For small scales, where $\xi > 1$, one should consider using a larger random catalog and one can use either the standard LS method or the split method with a more modest split factor. Because the

number of pairs with these small separations is much smaller, the computation time is not a similar issue as for large separations.

The results of our analysis will have an impact also on the computationally more demanding task of covariance matrix estimation. However, since in that case the exact computational cost is determined by the balance of data catalogs and random catalogs, which does not need to be the same as for the individual two-point correlation estimate, we postpone a quantitative analysis to a future, dedicated study. The same kind of methods can be applied to higher-order statistics (three-point and four-point correlation functions) to speed up their estimation (Slepian & Eisenstein 2015).

*Acknowledgements.* We thank Will Percival and Cristiano Porciani for useful discussions. The 2PCF computations were done at the Euclid Science Data Center Finland (SDC-FI, urn:nbn:fi:research-infras-2016072529), for whose computational resources we thank CSC – IT Center for Science, the Finnish Grid and Cloud Computing Infrastructure (FGCI, urn:nbn:fi:research-infras-2016072533), and the Academy of Finland grant 292882. This work was supported by the Academy of Finland grant 295113. VL was supported by the Jenny and Antti Wihuri Foundation, AV by the Väisälä Foundation, AS by the Magnus Ehrnrooth Foundation, and JV by the Finnish Cultural Foundation. We also acknowledge travel support from the Jenny and Antti Wihuri Foundation. VA acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 749348. FM acknowledges the grants ASI n.I/023/12/0 "Attivit'a relative alla fase B2/C per la missione Euclid" and PRIN MIUR 2015 "Cosmology and Fundamental Physics: illuminating the Dark Universe with Euclid".

## References

Alam, S., Ata, M., Bailey, S., et al. 2017, MNRAS, 470, 2617
Alcock, C., & Paczyński, B. 1979, Nature, 281, 358
Alonso, D. 2012, ArXiv e-prints [arXiv:1210.1833]
Anderson, L., Aubourg, E., Bailey, S., et al. 2012, MNRAS, 427, 3435
Anderson, L., Aubourg, E., Bailey, S., et al. 2014, MNRAS, 441, 24
Ata, M., Baumgarten, F., Bautista, J., et al. 2018, MNRAS, 473, 4773
Bautista, J. E., Vargas-Magaña, M., Dawson, K. S., et al. 2018, ApJ, 863, 110
Bernstein, G. M. 1994, ApJ, 424, 569
Beutler, F., Blake, C., Colless, M., et al. 2011, MNRAS, 416, 3017
Beutler, F., Blake, C., Colless, M., et al. 2012, MNRAS, 423, 3430
Blake, C., Kazin, E. A., Beutler, F., et al. 2011, MNRAS, 418, 1707
Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505
Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465
de la Torre, S., Jullo, E., Giocoli, C., et al. 2017, A&A, 608, A44
Demina, R., Cheong, S., BenZvi, S., & Hindrichs, O. 2018, MNRAS, 480, 49
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560
Grieb, J. N., Sánchez, A. G., Salvador-Albornoz, S., & Dalla Vecchia, C. 2016, MNRAS, 457, 1577
Guzzo, L., Pierleoni, M., Meneux, B., et al. 2008, Nature, 451, 541
Hamilton, A. J. S. 1993, ApJ, 417, 19
Hewett, H. C. 1982, MNRAS, 201, 867
Hou, J., Sánchez, A. G., Scoccimarro, R., et al. 2018, MNRAS, 480, 2521
Jarvis, M. 2015, Astrophysics Source Code Library [record ascl:1508.007]
Kaiser, N. 1987, MNRAS, 227, 1
Kerscher, M. 1999, A&A, 343, 333
Kerscher, M., Szapudi, I., & Szalay, A. S. 2000, ApJ, 535, L13
Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Lippich, M., Sánchez, A. G., Colavincenzo, M., et al. 2019, MNRAS, 482, 1786
Marulli, F., Veropalumbo, A., & Moresco, M. 2016, Astron. Comput., 14, 35
Moore, A., Connolly, A., Genovese, C., et al. 2000, ArXiv e-prints [arXiv:astro-ph/0012333]
Peacock, J. A., Cole, S., Norberg, P., et al. 2001, Nature, 410, 169
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
Pezzotta, A., de la Torre, S., Bel, J., et al. 2017, A&A, 604, A33
Reid, B. A., Samushia, L., White, M., et al. 2012, MNRAS, 426, 2719
Ross, A. J., Samushia, L., Howlett, C., et al. 2015, MNRAS, 449, 835
Ross, A. J., Beutler, F., Chuang, C. H., et al. 2017, MNRAS, 464, 1168
Ruggeri, R., Percival, W. J., Gil-Marín, H., et al. 2019, MNRAS, 483, 3878
Slepian, Z., & Eisenstein, D. J. 2015, MNRAS, 454, 4142
Vargas-Magaña, M., Ho, S., Cuesta, A. J., et al. 2018, MNRAS, 477, 1153
Wall, J. V., & Jenkins, C. R. 2012, Practical Statistics for Astronomers (Cambridge, UK: Cambridge University Press)
Zarrouk, P., Burtin, E., Gil-Marín, H., et al. 2018, MNRAS, 477, 1639
Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, ApJ, 736, 59

## Appendix A: Derivation examples

As examples of how the variances of the different deviations, $\langle\alpha\beta\rangle$ and so on, are derived, following the method presented in LS93, we give three of the cases here: $\langle\alpha_1\alpha_2\rangle$ (common for all the variants of LS), and $\langle\beta_1\gamma'_2\rangle$ and $\langle\gamma'_1\gamma'_2\rangle$ for the split method. The rest are calculated in a similar manner.

To derive the $\langle DD_1\ DD_2\rangle$ appearing in the $\langle\alpha_1\alpha_2\rangle$ of (15) we start from

$$\langle DD_1\ DD_2\rangle = \sum_{i<j}^{K}\sum_{k<l}^{K}\langle n_i n_j n_k n_l\rangle\Theta^{ij}(\boldsymbol{r}_1)\Theta^{kl}(\boldsymbol{r}_2), \tag{A.1}$$

where both $i, j$, and $k, l$ sum over all microcell pairs; $n_i = 1$ or 0 is the number of galaxies in microcell $i$. There are three different cases for the terms $\langle n_i n_j n_k n_l\rangle$ depending on how many indices are equal ($i \neq j$ and $k \neq l$ for all of them).

The first case (quadruplets, i.e., two pairs of microcells) is when $i, j, k, l$ are all different. We use $\sum^*_{ijkl}$ to denote this part of the sum. There are $\frac{1}{2}K(K-1)\times\frac{1}{2}(K-2)(K-3) = \frac{1}{4}K^4$ (we work in the limit $K \to \infty$) such terms and they have

$$\begin{aligned}
&\langle n_i n_j n_k n_l\rangle \\
&\quad= \frac{N_\mathrm{d}}{K}\frac{N_\mathrm{d}-1}{K-1}\frac{N_\mathrm{d}-2}{K-2}\frac{N_\mathrm{d}-3}{K-3}\Big\langle(1+\delta_i)(1+\delta_j)(1+\delta_k)(1+\delta_l)\Big\rangle \\
&\quad= \frac{N_\mathrm{d}(N_\mathrm{d}-1)(N_\mathrm{d}-2)(N_\mathrm{d}-3)}{K^4}\Big[1+\langle\delta_i\delta_j\rangle+\langle\delta_k\delta_l\rangle+\langle\delta_i\delta_k\rangle \\
&\qquad+\langle\delta_i\delta_l\rangle+\langle\delta_j\delta_k\rangle+\langle\delta_j\delta_l\rangle \\
&\qquad+\langle\delta_i\delta_j\delta_k\rangle+\langle\delta_i\delta_j\delta_l\rangle+\langle\delta_i\delta_k\delta_l\rangle+\langle\delta_j\delta_k\delta_l\rangle+\langle\delta_i\delta_j\delta_k\delta_l\rangle\Big] \\
&\quad= \frac{N_\mathrm{d}(N_\mathrm{d}-1)(N_\mathrm{d}-2)(N_\mathrm{d}-3)}{K^4}\Big[1+\xi(\boldsymbol{r}_{ij})+\xi(\boldsymbol{r}_{kl})+\xi(\boldsymbol{r}_{ik}) \\
&\qquad+\xi(\boldsymbol{r}_{il})+\xi(\boldsymbol{r}_{jk})+\xi(\boldsymbol{r}_{jl}) \\
&\qquad+\zeta(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)+\zeta(\boldsymbol{x}_i,\boldsymbol{x}_k,\boldsymbol{x}_l)+\zeta(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_l)+\zeta(\boldsymbol{x}_j,\boldsymbol{x}_k,\boldsymbol{x}_l) \\
&\qquad+\xi(\boldsymbol{r}_{ij})\xi(\boldsymbol{r}_{kl})+\xi(\boldsymbol{r}_{ik})\xi(\boldsymbol{r}_{jl})+\xi(\boldsymbol{r}_{il})\xi(\boldsymbol{r}_{jk}) \\
&\qquad+\eta(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k,\boldsymbol{x}_l)\Big],
\end{aligned} \tag{A.2}$$

where $\delta_i := \delta(\boldsymbol{x}_i)$ is the density perturbation (relative to the actual mean density of galaxies in the survey), $\zeta$ is the three-point correlation, and $\eta$ is the connected (i.e., the part that does not arise from the two-point correlation) four-point correlation. The fraction of microcell quadruplets (pairs of pairs) that satisfy $\boldsymbol{r}_{ij} \equiv \boldsymbol{x}_j - \boldsymbol{x}_i \in \boldsymbol{r}_1$ and $\boldsymbol{r}_{kl} \in \boldsymbol{r}_2$ is $G^\mathrm{p}(\boldsymbol{r}_1)G^\mathrm{p}(\boldsymbol{r}_2) =: G^\mathrm{p}_1 G^\mathrm{p}_2$. In the limit of large $K$ the number of other index quadruplets is negligible compared to those where all indices have different values, so we have

$$\sum_{ijkl}^{*}\Theta^{ij}_1\Theta^{kl}_2 = \frac{K(K-1)(K-2)(K-3)}{4}G^\mathrm{p}_1 G^\mathrm{p}_2 = \frac{K^4}{4}G^\mathrm{p}_1 G^\mathrm{p}_2. \tag{A.3}$$

For the connected pairs $(i, j)$ and $(k, l)$ we have $\xi(\boldsymbol{r}_{ij}) = \xi(\boldsymbol{r}_1) \equiv \xi_1$ and $\xi(\boldsymbol{r}_{kl}) = \xi(\boldsymbol{r}_2) \equiv \xi_2$.

The second case (triplets of microcells) is when $k$ or $l$ is equal to $i$ or $j$. We denote this part of the sum by

$$\sum_{ijk}^{*}\langle n_i n_j n_k\rangle\Theta^{ik}_1\Theta^{jk}_2. \tag{A.4}$$

It turns out that it goes over all ordered combinations of $\{i, j, k\}$, where $i, j, k$ are all different, exactly once, so there are $K(K-1)(K-2) = K^3$ such terms (triplets). Here

$$\begin{aligned}
\langle n_i n_j n_k\rangle &= \frac{N_\mathrm{d}}{K}\frac{N_\mathrm{d}-1}{K-1}\frac{N_\mathrm{d}-2}{K-2}\Big\langle(1+\delta_i)(1+\delta_j)(1+\delta_k)\Big\rangle \\
&= \frac{N_\mathrm{d}(N_\mathrm{d}-1)(N_\mathrm{d}-2)}{K^3}\Big[1+\langle\delta_i\delta_k\rangle+\langle\delta_j\delta_k\rangle+\langle\delta_i\delta_j\rangle+\langle\delta_i\delta_j\delta_k\rangle\Big] \\
&= \frac{N_\mathrm{d}(N_\mathrm{d}-1)(N_\mathrm{d}-2)}{K^3}\Big[1+\xi(\boldsymbol{r}_{ik})+\xi(\boldsymbol{r}_{jk})+\xi(\boldsymbol{r}_{ij})+\zeta(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)\Big],
\end{aligned} \tag{A.5}$$

and

$$\sum_{ijk}^{*}\Theta^{ik}_1\Theta^{jk}_2 = K^3 G^\mathrm{t}_{12}. \tag{A.6}$$

The third case (pairs of microcells) is when $i = k$ and $j = l$. This part of the sum becomes

$$\sum_{i<j}\langle n_i n_j\rangle\Theta^{ij}_1\Theta^{ij}_2, \tag{A.7}$$

where

$$\langle n_i n_j\rangle = \frac{N_\mathrm{d}(N_\mathrm{d}-1)}{K^2}\Big[1+\xi(\boldsymbol{r}_{ij})\Big], \tag{A.8}$$

and

$$\sum_{i<j}\Theta^{ij}_1\Theta^{ij}_2 = \delta_{12}\frac{K^2}{2}G^\mathrm{p}_1, \tag{A.9}$$

that is, the sum vanishes unless the two bins are the same, $\boldsymbol{r}_1 = \boldsymbol{r}_2$.

We now apply the three approximations listed in Sect. 2.2: (1) $\xi(\boldsymbol{r}_{ik}) = \xi(\boldsymbol{r}_{il}) = \xi(\boldsymbol{r}_{jk}) = \xi(\boldsymbol{r}_{jl}) = 0$ in (A.2); (2) $\zeta = 0$ in (A.2) and (A.5); (3) $\eta = 0$ in (A.2). We obtain

$$\begin{aligned}
\langle DD_1\ DD_2\rangle \simeq\ &\frac{1}{4}N_\mathrm{d}(N_\mathrm{d}-1)(N_\mathrm{d}-2)(N_\mathrm{d}-3)(1+\xi_1)(1+\xi_2)G^\mathrm{p}_1 G^\mathrm{p}_2 \\
&+N_\mathrm{d}(N_\mathrm{d}-1)(N_\mathrm{d}-2)(1+\xi_1+\xi_2+\xi_{12})G^\mathrm{t}_{12} \\
&+\frac{1}{2}N_\mathrm{d}(N_\mathrm{d}-1)\delta_{12}(1+\xi_1)G^\mathrm{p}_1,
\end{aligned} \tag{A.10}$$

and using $\langle DD\rangle$ from (3) we arrive at the $\langle\alpha_1\alpha_2\rangle$ result given in Eq. (15).

For the split method, we give the calculation for

$$\begin{aligned}
\langle\gamma'_1\beta'_2\rangle &= \frac{\langle RR'_1\ DR_2\rangle - \langle RR'_1\rangle\langle DR_2\rangle}{\langle RR'_1\rangle\langle DR_2\rangle}, \\
\langle\gamma'_1\gamma'_2\rangle &= \frac{\langle RR'_1\ RR'_2\rangle - \langle RR'_1\rangle\langle RR'_2\rangle}{\langle RR'_1\rangle\langle RR'_2\rangle}
\end{aligned} \tag{A.11}$$

below.

First the $\langle\gamma'_1\beta_2\rangle$:

$$\langle RR'_1\ DR_2\rangle = \sum_{\mu=1}^{M_\mathrm{s}}\langle R^\mu R^\mu_1\ DR_2\rangle = M_\mathrm{s}\langle R^\mu R^\mu_1\ DR_2\rangle, \tag{A.12}$$

where

$$\langle R^\mu R^\mu_1\ DR_2\rangle = \sum_{i<j}\sum_{k\neq l}\langle s_i s_j n_k r_l\rangle\Theta^{ij}_1\Theta^{kl}_2, \tag{A.13}$$

$s_i$ is the number (0 or 1) of $R^\mu$ objects in microcell $i$, and $r_l$ is the number of $R$ objects in microcell $l$.

There are $\frac{1}{2}K^4$ quadruplet terms, for which (see (A.3)) $\sum^* \Theta\Theta = \frac{1}{2}K^4 G_1^{\rm p} G_2^{\rm p}$, and

$$\langle s_i s_j n_k r_l \rangle = \frac{N_{\rm r}'(N_{\rm r}'-1)N_{\rm d}(N_{\rm r}'-2)}{K^4}. \tag{A.14}$$

Triplets where $i$ or $j$ is equal to $k$ have $\langle s_i s_k n_k r_l \rangle = 0$, since $s_k n_k = 0$ always (we cannot have two objects, from $R^\mu$ and $D$, in the same cell). There are $K^3$ triplet terms where $i$ or $j$ is equal to $l$: for them $\sum^* \Theta\Theta = K^3 G_{12}^{\rm t}$, and

$$\langle s_i s_l n_k r_l \rangle = \langle s_i s_l n_k \rangle = \frac{N_{\rm r}'(N_{\rm r}'-1)N_{\rm d}}{K^3}, \tag{A.15}$$

where if $s_l = 1$ then also $r_l = 1$ since $R^\mu \subset R$.

Pairs where $(i, j) = (k, l)$ or $(i, j) = (l, k)$ have $\langle s_k s_l n_k r_l \rangle = 0$, since again we cannot have two different objects in the same cell. Thus

$$\langle RR_1' \; DR_2 \rangle = \frac{1}{2} N_{\rm d} N_{\rm r}(N_{\rm r}'-1)(N_{\rm r}'-2) G_1^{\rm p} G_2^{\rm p} + N_{\rm d} N_{\rm r}(N_{\rm r}'-1) G_{12}^{\rm t}, \tag{A.16}$$

and we obtain that

$$\langle \gamma_1' \beta_2 \rangle = \frac{2}{N_{\rm r}'} \left[ \frac{G_{12}^{\rm t}}{G_1^{\rm p} G_2^{\rm p}} - 1 \right] = 2\, t_{12}^{\rm r}, \tag{A.17}$$

which is equal to the $\langle \gamma_1 \beta_2 \rangle$ of standard LS.

Then the $\langle \gamma_1' \gamma_2' \rangle$:

$$\langle RR_1' \; RR_2' \rangle = \sum_{\mu=1}^{M_{\rm s}} \sum_{\nu=1}^{M_{\rm s}} \langle R^\mu R_1^\mu \cdot R^\nu R_2^\nu \rangle, \tag{A.18}$$

where there are $M_{\rm s}(M_{\rm s}-1)$ terms with $\mu \neq \nu$ giving

$$\langle R^\mu R_1^\mu \; R^\nu R_2^\nu \rangle = \langle R^\mu R_1^\mu \rangle \langle R^\nu R_2^\nu \rangle = \frac{1}{4}(N_{\rm r}')^2 (N_{\rm r}'-1)^2 G_1^{\rm p} G_2^{\rm p}, \tag{A.19}$$

and $M_{\rm s}$ terms with $\mu = \nu$ giving

$$\begin{aligned} \langle R^\mu R_1^\mu \; R^\mu R_2^\mu \rangle = & \frac{1}{4} N_{\rm r}'(N_{\rm r}'-1)(N_{\rm r}'-2)(N_{\rm r}'-3) G_1^{\rm p} G_2^{\rm p} \\ & + N_{\rm r}'(N_{\rm r}'-1)(N_{\rm r}'-2) G_{12}^{\rm t} \\ & + \frac{1}{2} N_{\rm r}'(N_{\rm r}'-1) \delta_{12} G_1^{\rm p}. \end{aligned} \tag{A.20}$$

Adding these up gives

$$\begin{aligned} \langle RR_1' \; RR_2' \rangle = & \frac{1}{4} N_{\rm r}'(N_{\rm r}'-1)(N_{\rm r}' N_{\rm r}' - N_{\rm r}' - 4\, N_{\rm r}' + 6) G_1^{\rm p} G_2^{\rm p} \\ & + N_{\rm r}'(N_{\rm r}'-1)(N_{\rm r}'-2) G_{12}^{\rm t} \\ & + \frac{1}{2} N_{\rm r}'(N_{\rm r}'-1) \delta_{12} G_1^{\rm p} \end{aligned} \tag{A.21}$$

and

$$\begin{aligned} \langle \gamma_1' \gamma_2' \rangle = & \frac{2}{N_{\rm r}'(N_{\rm r}'-1)} \left\{ 2(N_{\rm r}'-2) \left[ \frac{G_{12}^{\rm t}}{G_1^{\rm p} G_2^{\rm p}} - 1 \right] + \frac{\delta_{12}}{G_1^{\rm p}} - 1 \right\} \\ = & \; 4\, t_{12}^{\rm r} + p_{12}^{\rm r'}. \end{aligned} \tag{A.22}$$

This differs both from standard LS and from dilution, since it involves both $N_{\rm r}'$ and $N_{\rm r}'$.