Euclid: fast two-point correlation function covariance through linear construction

Keihänen, E.; Lindholm, V.; Monaco, P.; Blot, L.; Carbone, C.; Kiiveri, K.; ... ; de la Torre, S.

A&A 666, A129 (2022)
https://doi.org/10.1051/0004-6361/202244065

Astronomy & Astrophysics

Euclid: Fast two-point correlation function covariance through linear construction⋆

E. Keihänen[1], V. Lindholm[2], P. Monaco[3,4,5,6], L. Blot[7], C. Carbone[8], K. Kiiveri[2], A. G. Sánchez[9], A. Viitanen[2], J. Valiviita[10], A. Amara[11], N. Auricchio[12], M. Baldi[13,12,14], D. Bonino[15], E. Branchini[16,17], M. Brescia[18], J. Brinchmann[19], S. Camera[20,21,15], V. Capobianco[15], J. Carretero[22,23], M. Castellano[24], S. Cavuoti[18,25,26], A. Cimatti[27,28], R. Cledassou[29,30], G. Congedo[31], L. Conversi[32,33], Y. Copin[34], L. Corcione[15], M. Cropper[35], A. Da Silva[36,37], H. Degaudenzi[38], M. Douspis[39], F. Dubath[38], C. A. J. Duncan[40], X. Dupac[33], S. Dusini[41], A. Ealet[34], S. Farrens[42], S. Ferriol[34], M. Frailis[5], E. Franceschi[12], M. Fumana[8], B. Gillis[31], C. Giocoli[43,44], A. Grazian[45], F. Grupp[9,46], L. Guzzo[47,48,49], S. V. H. Haugan[50], H. Hoekstra[51], W. Holmes[52], F. Hormuth[53], K. Jahnke[54], M. Kümmel[46], S. Kermiche[55], A. Kiessling[52], T. Kitching[35], M. Kunz[56], H. Kurki-Suonio[1,10], S. Ligori[15], P. B. Lilje[50], I. Lloro[57], E. Maiorano[12], O. Mansutti[5], O. Marggraf[58], F. Marulli[13,12,14], R. Massey[59], M. Melchior[60], M. Meneghetti[12,61], G. Meylan[62], M. Moresco[13,12], B. Morin[42], L. Moscardini[13,12,14], E. Munari[5], S. M. Niemi[63], C. Padilla[22], S. Paltani[38], F. Pasian[5], K. Pedersen[64], V. Pettorino[65], S. Pires[42], G. Polenta[66], M. Poncet[29], L. Popa[67], F. Raison[9], A. Renzi[68,41], J. Rhodes[52], E. Romelli[5], R. Saglia[9,46], B. Sartoris[5,4], P. Schneider[58], T. Schrabback[58], A. Secroun[55], G. Seidel[54], C. Sirignano[68,41], G. Sirri[14], L. Stanco[41], C. Surace[69], P. Tallada-Crespí[70,23], D. Tavagnacco[5], A. N. Taylor[31], I. Tereno[36,71], R. Toledo-Moreo[72], F. Torradeflot[23,70], E. A. Valentijn[73], L. Valenziano[12,14], T. Vassallo[5], Y. Wang[74], J. Weller[9,46], G. Zamorani[12], J. Zoubian[55], S. Andreon[48], D. Maino[47,8,49], and S. de la Torre[69]

(Affiliations can be found after the references)

Received 20 May 2022 / Accepted 27 June 2022

ABSTRACT

We present a method for fast evaluation of the covariance matrix for a two-point galaxy correlation function (2PCF) measured with the Landy–Szalay estimator. The standard way of evaluating the covariance matrix consists in running the estimator on a large number of mock catalogs, and evaluating their sample covariance. With large random catalog sizes (random-to-data objects' ratio $M \gg 1$) the computational cost of the standard method is dominated by that of counting the data-random and random-random pairs, while the uncertainty of the estimate is dominated by that of data-data pairs. We present a method called Linear Construction (LC), where the covariance is estimated for small random catalogs with a size of $M = 1$ and $M = 2$, and the covariance for arbitrary M is constructed as a linear combination of the two. We show that the LC covariance estimate is unbiased. We validated the method with PINOCCHIO simulations in the range $r = 20$–$200\,h^{-1}$ Mpc. With $M = 50$ and with $2\,h^{-1}$ Mpc bins, the theoretical speedup of the method is a factor of 14. We discuss the impact on the precision matrix and parameter estimation, and present a formula for the covariance of covariance.

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

1. Introduction

The next generation of telescopes for cosmology surveys, such as *Euclid* (Laureijs et al. 2011), the *Vera Rubin* Observatory (Ivezić et al. 2019), DESI (DESI Collaboration 2016), or the *Nancy Grace Roman* Space Telescope (Akeson et al. 2019), will soon greatly improve the quality and quantity of data for galaxy clustering and lensing measurements. Their main aim is to illuminate the dark sector of cosmology, to test Einstein's gravity on large scales, and to find signatures of the physics of inflation such as primordial non-Gaussianities.

Galaxy clustering is one of our most powerful cosmological probes (Cole et al. 2005; Eisenstein et al. 2005; Alam et al. 2005; Alam et al. 2021). However, galaxies are biased tracers of the nonlinear density field and their selection is subject to several different effects, such as fluctuations in exposure time, noise level, Milky Way extinction, photometry calibration error, sample contamination among others (e.g., Jasche & Lavaux 2017; Monaco et al. 2019; Kalus et al. 2019; Merz et al. 2021). Thus, their clustering will contain entangled information of matter clustering, galaxy bias, and observational systematics. The uncertainty will be represented by a covariance matrix. The inverse of the covariance matrix, the precision matrix, will be used in the likelihood analysis to infer cosmological parameters and their covariance. An accurate quantification of the clustering covariance under all the sources of uncertainty is therefore of paramount importance for the success of a survey.

It is customary to construct covariance matrices of galaxy clustering by using large samples of hundreds, if not thousands, of mock galaxy catalogs in the past light cone (e.g., Manera et al. 2013; Kitaura et al. 2016). Any known selection effects are applied to the mock catalogs, after which the clustering signal is measured with the same procedure as the one used for the actual data catalog. This sample of clustering

⋆ This paper is published on behalf of the *Euclid* Consortium.

measurements makes it possible to construct a brute-force numerical sample covariance. This approach has many advantages. It is conceptually simple, and the covariance built this way is positive-definite by construction. The estimation error of the covariance and its propagation to parameter estimation are well understood (Taylor et al. 2013; Taylor & Joachimi 2014; Hartlap et al. 2007; Dodelson & Schneider 2013; Percival et al. 2014, 2022; Sellentin & Heavens 2016). The sample covariance, however, is computationally expensive to construct; the variance of the covariance estimate decreases proportionally to $1/N$, where N is the number of simulations, so getting the error down to a 10% level requires about 100 independent realizations, and ∼10 000 realizations for a 1% level.

This raises two related problems: on the one hand, the production of such mocks, which are typically addressed with approximate methods to bypass the high cost of N-body simulations (Monaco 2016); on the other hand, the measurement of galaxy clustering of thousands of mocks, which can be a bottleneck for a processing pipeline. Several strategies have been proposed to reduce the cost of covariance estimation. These include precision matrix expansion (Friedrich & Eifler 2018), tapering (Paz & Sánchez 2015), eigenvalue decomposition (Gaztañaga & Scoccimarro 2005), linear shrinkage (Pope & Szapudi 2008), sparse precision matrix estimation (Padmanabhan et al. 2016), and nonlinear shrinkage (Joachimi 2017).

In this work we focus on the estimation of the two-point correlation function (2PCF) and its covariance. In the special case of Gaussian fluctuations, the 2PCF contains all information on the statistical properties of the galaxy distribution. A concrete example would be the European Space Agency's *Euclid* cosmology mission, and in particular its spectroscopic sample of Hα emitting galaxies (Euclid Collaboration 2022). This galaxy sample is expected to be as large as 20–30 million galaxies in the redshift range 0.9–1.8. The Euclid Consortium plans to represent the covariance matrix of the 2PCF with a few thousand mock catalogs. The time needed to measure the 2PCF of such a large number of mocks will be one major contributor to the whole pipeline from raw images to cosmological parameter inference.

The Landy–Szalay estimator (Landy & Szalay 1993) has become the standard estimator in galaxy clustering science. In addition to the actual galaxy catalog, the Landy–Szalay estimator requires a random catalog, which represents a uniform distribution of points within the survey volume considered, modulated with same weighting and selection as the data catalog. The 2PCF is then built as a combination of data-data (DD), data-random (DR), and random-random (RR) pair counts. The estimator is unbiased at the limit of an infinite random catalog, and, when the fluctuations are small, it yields the minimum-variance solution for the correlation function.

Since the random catalog is usually much larger (in number of objects) than the data catalog, the computational cost of the estimator is dominated by the cost of the RR counts. Glass-like random catalogs (Dávila-Kurbán et al. 2021) have been proposed as a way of reducing the required random catalog size. A straightforward way to reduce the cost is to coadd RR pair counts from a collection of small subcatalogs, rather than counting the pairs in one large catalog, thus omitting pairs between subcatalogs. This natural idea has been applied in many studies without explicit mention, or without assigning a name to it. We refer to this approach as the "split" method, thusly named because of the idea of "splitting" a large random catalog into several small ones. The term was coined in Keihänen et al. (2019), where the effect of the size of the random catalog on the estimator error is studied in a systematic way, and it is shown that the effect of the splitting on the estimation error is negligible. It is also shown that the optimal relation between the accuracy and computational cost is achieved when the subcatalogs have the same number of objects as the data catalog.

Even with a split random catalog, most of the computation time goes into the counting of the RR and DR pairs, while the estimation error is dominated by the scatter of the data points. The same applies to the sampling of the covariance matrix, the cost of which is N times that of a single 2PCF estimation. Using a single random catalog for all measurements can reduce the cost of the RR counts, but then counting the DR pairs becomes the next bottleneck.

In this paper we introduce a way of speeding up the covariance estimation, specific to the Landy–Szalay estimator. We aim to show that the covariance matrix for a 2PCF estimate can be constructed using a significantly smaller random catalog than what was used in the construction of the 2PCF itself.

The paper is organized as follows. In Sect. 2 we present the method and its theoretical background. In Sect. 3 we discuss the accuracy of the method, derive a covariance of covariance, and discuss implications for parameter estimation. In Sect. 4 we describe the simulations we used for the validation of the method. In Sect. 5 we present our results, comparing the accuracy and speed of the new method to those of the sample covariance. We give our conclusions in Sect. 6.

This work has made use of the 2PCF code developed by the Euclid Consortium.

2. Method

2.1. Landy–Szalay estimator

We denote the number of objects in the data catalog by N_d, and in the random catalog by N_r. We assume that the two-point correlation function is estimated with the Landy–Szalay estimator, with the additional twist of the "split" option, as follows. The random catalog is split into M subcatalogs, where M is called the split factor. RR pairs are counted within each subcatalog and coadded, but pairs of objects in two distinct subcatalogs are omitted. Each subcatalog will have to obey the same statistical properties and have the same sky coverage as the full catalog. In other words, each subcatalog must itself be a valid random catalog. Splitting the random catalog reduces the computational cost of 2PCF estimation significantly, for a negligible loss of accuracy (Keihänen et al. 2019). The optimal split factor has been shown to be $M = N_\mathrm{r}/N_\mathrm{d}$, that is to say the random catalog is split into subcatalogs of the same size as the data catalog. For fixed N_r, this minimizes the variance of the correlation function for a given computation time, or minimizes the computation time required to reach a given target variance. From here on we parametrize the size of the random catalog as $N_\mathrm{r} = MN_\mathrm{d}$.

The random catalog is usually constructed to be significantly larger than the data catalog, in order that the estimation error is dominated by the scatter of the data points rather than that of the random points. In this work we use as baseline the value $M = 50$, the value adopted for the *Euclid* galaxy clustering study.

The Landy–Szalay estimator is

$$\hat{\xi}(\mathbf{r}) := \frac{\mathrm{dd}(\mathbf{r}) - 2\mathrm{dr}(\mathbf{r})}{\mathrm{rr}(\mathbf{r})} + 1, \tag{1}$$

where $\mathrm{dd}(\mathbf{r})$, $\mathrm{dr}(\mathbf{r})$, and $\mathrm{rr}(\mathbf{r})$ denote the normalized data-data, data-random, and random-random pair counts in a separation bin. Following the notation of Keihänen et al. (2019), we use the

vector $\mathbf{r}$ to denote the generalized separation vector. Vector $\mathbf{r}$ may refer to a physical separation distance, or, more generally, to an arbitrary bin in 1D, 2D, or 3D space. The normalized data-data pair count is

$$\mathrm{dd}(\mathbf{r}) := \frac{\mathrm{DD}(\mathbf{r})}{N_\mathrm{d}(N_\mathrm{d}-1)/2}, \tag{2}$$

where $\mathrm{DD}(\mathbf{r})$ is the unnormalized pair count. This is unaffected by the split. Similarly, the normalized data-random count is given by

$$\mathrm{dr}(\mathbf{r}) := M^{-1}\sum_{i=1}^{M}\frac{\mathrm{DR}_i(\mathbf{r})}{N_\mathrm{d}^2} = M^{-1}\sum_{i=1}^{M}\mathrm{dr}_i(\mathbf{r}), \tag{3}$$

where DR_i is the pair count between the data catalog and the ith random subcatalog. Since the dependence on the random catalog is linear, this too is unaffected by the split.

The normalized random-random count with split can be written as

$$\mathrm{rr}(\mathbf{r}) := M^{-1}\sum_{i=1}^{M}\frac{\mathrm{RR}_i(\mathbf{r})}{N_\mathrm{d}(N_\mathrm{d}-1)/2} = M^{-1}\sum_{i=1}^{M}\mathrm{rr}_i(\mathbf{r}), \tag{4}$$

where RR_i is the unnormalized pair count from the ith subcatalog. With this notation, the split Landy–Szalay estimator becomes

$$\hat{\xi}(\mathbf{r}) = \frac{\mathrm{dd}(\mathbf{r}) - \frac{2}{M}\sum_i \mathrm{dr}_i(\mathbf{r})}{M^{-1}\sum_i \mathrm{rr}_i(\mathbf{r})} + 1. \tag{5}$$

We use the hat ($\hat{\xi}$) to indicate that this is an estimate of the true correlation function ξ.

The computational cost of the estimator is roughly proportional to the total number of pairs counted. The cost of the split estimator is proportional to $\frac{1}{2}N_\mathrm{d}^2(1+3M)$, while without split the cost grows proportional to $\frac{1}{2}N_\mathrm{d}^2(1+2M+M^2)$.

2.2. Covariance

We consider the estimated correlation function in two distance bins $\mathbf{r}_1$ and $\mathbf{r}_2$, which may or may not be the same. We want to estimate the covariance

$$\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2)\right] := \left\langle\left[\hat{\xi}(\mathbf{r}_1)-\langle\hat{\xi}(\mathbf{r}_1)\rangle\right]\left[\hat{\xi}(\mathbf{r}_2)-\langle\hat{\xi}(\mathbf{r}_2)\rangle\right]\right\rangle. \tag{6}$$

The brackets $\langle\rangle$ denote an average over an infinite ensemble of data realizations, for fixed cosmology and survey geometry. Since we consider the actual measured correlation function to represent one such realization, the covariance is a measure of the statistical uncertainty in the measured correlation function.

Assume we have N mock catalogs and corresponding random catalogs, with the same sky coverage, masking etc. as the actual survey catalog. Let $\hat{\xi}_i(\mathbf{r}), i = 1\ldots N$ denote the set of correlation functions estimated from these mocks. An unbiased estimate of the covariance is given by the sample covariance

$$\hat{C}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2)\right] := \frac{1}{N-1}\sum_{i=1}^{N}\left[\hat{\xi}_i(\mathbf{r}_1)-\bar{\xi}(\mathbf{r}_1)\right]\left[\hat{\xi}_i(\mathbf{r}_2)-\bar{\xi}(\mathbf{r}_2)\right], \tag{7}$$

where

$$\bar{\xi}(\mathbf{r}) := \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i(\mathbf{r}) \tag{8}$$

is the estimated mean. The required number of mocks will depend on the accuracy requirement of the application in question. For few percent-level accuracy in parameter error bars, $N \gtrsim 1000$ is required.

Throughout this paper we use a convention where the symbol $\hat{\mathrm{C}}$ with a hat denotes a numerical estimate of a covariance, and either cov or C denotes the true (ensemble average) covariance. Specifically, $\hat{C}$ is reserved for the sample covariance estimate, constructed as in Eq. (7) (see Table 1).

The computational cost of constructing the sample covariance, obviously, is N times the cost of a single 2PCF estimate. In this paper we show that a given level of accuracy can be reached with a significantly lower CPU cost. For this goal, we now break the covariance of the Landy–Szalay estimator into pair count covariances.

Following the notation of the Landy–Szalay paper, we write

$$\begin{aligned}\mathrm{dd}(\mathbf{r}) &= \langle\mathrm{dd}(\mathbf{r})\rangle\left[1+\alpha(\mathbf{r})\right],\\ \mathrm{dr}_i(\mathbf{r}) &= \langle\mathrm{dr}(\mathbf{r})\rangle\left[1+\beta_i(\mathbf{r})\right],\\ \mathrm{rr}_i(\mathbf{r}) &= \langle\mathrm{rr}(\mathbf{r})\rangle\left[1+\gamma_i(\mathbf{r})\right].\end{aligned} \tag{9}$$

The brackets $\langle\rangle$ indicate an ensemble average. Thus α,β_i,γ_i capture the variation of the pair counts around their average. By definition, $\langle\alpha\rangle = \langle\beta\rangle = \langle\gamma\rangle = 0$. Inserting these into the Landy–Szalay estimator yields

$$\begin{aligned}\hat{\xi}(\mathbf{r}) = &\frac{\langle\mathrm{dd}(\mathbf{r})\rangle\left[1+\alpha(\mathbf{r})\right]}{\langle\mathrm{rr}(\mathbf{r})\rangle\left[1+M^{-1}\sum_i\gamma_i(\mathbf{r})\right]}\\ &-2\frac{\langle\mathrm{dr}(\mathbf{r})\rangle\left[1+M^{-1}\sum_i\beta_i(\mathbf{r})\right]}{\langle\mathrm{rr}(\mathbf{r})\rangle\left[1+M^{-1}\sum_i\gamma_i(\mathbf{r})\right]}+1.\end{aligned} \tag{10}$$

For the ensemble averages it holds that $\langle\mathrm{dr}(\mathbf{r})\rangle = \langle\mathrm{rr}(\mathbf{r})\rangle$ and $\langle\mathrm{dd}(\mathbf{r})\rangle = d(\mathbf{r})\langle\mathrm{rr}(\mathbf{r})\rangle$, where we define

$$d(\mathbf{r}) := 1+\xi(\mathbf{r}). \tag{11}$$

If the RR counts are large, as is usually the case, then $\gamma_i(\mathbf{r}) \ll 1$, and $[1+M^{-1}\sum_i\gamma_i(\mathbf{r})]^{-1} \approx 1-M^{-1}\sum_i\gamma_i(\mathbf{r})$. Assuming $\alpha,\beta,\gamma \ll 1$, we can drop the quadratic terms, and the estimator becomes

$$\begin{aligned}\hat{\xi}(\mathbf{r}) \approx\ & d(\mathbf{r})\left[1+\alpha(\mathbf{r})-M^{-1}\textstyle\sum_i\gamma_i(\mathbf{r})\right]\\ &-2\left[1+M^{-1}\textstyle\sum_i\beta_i(\mathbf{r})-M^{-1}\sum_i\gamma_i(\mathbf{r})\right]+1\end{aligned} \tag{12}$$

and as ensemble average $\langle\hat{\xi}(\mathbf{r})\rangle = d(\mathbf{r})-1$. The deviation from the ensemble average is

$$\begin{aligned}&\hat{\xi}(\mathbf{r})-\langle\xi(\mathbf{r})\rangle\\ &\quad\approx d(\mathbf{r})\left[\alpha(\mathbf{r})-M^{-1}\textstyle\sum_i\gamma_i(\mathbf{r})\right]-2M^{-1}\left[\sum_i\beta_i(\mathbf{r})-\sum_i\gamma_i(\mathbf{r})\right].\end{aligned} \tag{13}$$

Inserting Eq. (13) into Eq. (6) yields a combination of cross-correlation terms between α,β,γ. We now consider each of them in turn, and make use of our knowledge of their statistical properties to calculate the expectation values.

Let us begin with the term

$$M^{-2}\sum_{ij}\langle\gamma_i(\mathbf{r}_1)\gamma_j(\mathbf{r}_2)\rangle. \tag{14}$$

Indices i, j label independent random subcatalogs, all of which are statistically identical. We therefore have $\langle\gamma_i(\mathbf{r}_1)\gamma_j(\mathbf{r}_2)\rangle = 0$ for $i \neq j$, and $\langle\gamma_i(\mathbf{r}_1)\gamma_i(\mathbf{r}_2)\rangle = \langle\gamma(\mathbf{r}_1)\gamma(\mathbf{r}_2)\rangle$, where we drop the subscript to indicate an ensemble average that is the same for all subcatalogs. The covariance element becomes

$$M^{-2}\sum_{ij}\langle\gamma_i(\mathbf{r}_1)\gamma_j(\mathbf{r}_2)\rangle = M^{-1}\langle\gamma(\mathbf{r}_1)\gamma(\mathbf{r}_2)\rangle. \tag{15}$$

Based on similar arguments we can write

$$M^{-1}\sum_i \langle\alpha(\mathbf{r}_1)\beta_i(\mathbf{r}_2)\rangle = \langle\alpha(\mathbf{r}_1)\beta(\mathbf{r}_2)\rangle \tag{16}$$

and

$$M^{-1}\sum_i \langle\alpha(\mathbf{r}_1)\gamma_i(\mathbf{r}_2)\rangle = \langle\alpha(\mathbf{r}_1)\gamma(\mathbf{r}_2)\rangle\,. \tag{17}$$

Assuming that the random catalog and the data catalog are independent would allow us to drop the $\langle\alpha(\mathbf{r}_1)\gamma_i(\mathbf{r}_2)\rangle$ terms. This is however not necessarily always true. If the characteristics of the observed data catalog (mask, selection function) are used for the generation of the random catalog, a correlation may arise between the data catalog and the random catalog. Although such correlations are likely to be small, the assumption of independence is not relevant for the method we are developing, and we will thus not implement it.

When dealing with the terms involving β and γ, we split the sums into $i = j$ and $i \neq j$ parts, to obtain

$$\begin{aligned} &M^{-2}\sum_{ij}\langle\beta_i(\mathbf{r}_1)\beta_j(\mathbf{r}_2)\rangle \\ &\quad = M^{-1}\langle\beta(\mathbf{r}_1)\beta(\mathbf{r}_2)\rangle + \left(1-M^{-1}\right)\langle\beta(\mathbf{r}_1)\beta(\mathbf{r}_2)\rangle_{\rm cr}. \end{aligned} \tag{18}$$

Here the subscript cr ('cross') denotes that we are dealing with DR counts that involve two distinct random subcatalogs, however correlated through the shared data catalog.

Based on similar arguments, we obtain:

$$\begin{aligned} &M^{-2}\sum_{ij}\langle\beta_i(\mathbf{r}_1)\gamma_j(\mathbf{r}_2)\rangle \\ &\quad = M^{-1}\langle\beta(\mathbf{r}_1)\gamma(\mathbf{r}_2)\rangle + \left(1-M^{-1}\right)\langle\beta(\mathbf{r}_1)\gamma(\mathbf{r}_2)\rangle_{\rm cr}. \end{aligned} \tag{19}$$

As in the case of $\langle\alpha\gamma\rangle$, assuming independence between the random catalog and the data catalog would allow us to drop the second term, but this assumption is not relevant for the method under discussion.

We introduce a more concise notation, where we drop the arguments $\mathbf{r}_1$ and $\mathbf{r}_2$, and each term is interpreted as a symmetrized version of itself. When d is involved, it is paired with the first element. For instance, $d\langle\alpha\beta\rangle$ is to be read as

$$d\langle\alpha\beta\rangle = \frac{1}{2}\left[d(\mathbf{r}_1)\langle\alpha(\mathbf{r}_1)\beta(\mathbf{r}_2)\rangle + d(\mathbf{r}_2)\langle\alpha(\mathbf{r}_2)\beta(\mathbf{r}_1)\rangle\right], \tag{20}$$

similarly for the other pairs. In this notation, the covariance takes the form

$$\begin{aligned} &\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2);M\right] \\ &\quad = d^2\left[\langle\alpha\alpha\rangle + M^{-1}\langle\gamma\gamma\rangle - 2\langle\alpha\gamma\rangle\right] \\ &\qquad - 4d\left[\langle\alpha\beta\rangle - \langle\alpha\gamma\rangle - M^{-1}\langle\gamma\beta\rangle - (1-M^{-1})\langle\gamma\beta\rangle_{\rm cr} + M^{-1}\langle\gamma\gamma\rangle\right] \\ &\qquad + 4\Big[M^{-1}\langle\beta\beta\rangle + (1-M^{-1})\langle\beta\beta\rangle_{\rm cr} \\ &\qquad - 2M^{-1}\langle\beta\gamma\rangle - 2(1-M^{-1})\langle\beta\gamma\rangle_{\rm cr} + M^{-1}\langle\gamma\gamma\rangle\Big], \end{aligned} \tag{21}$$

where the third argument (M) indicates the size of the random catalog.

We have expressed the Landy–Szalay covariance in terms of pair-count covariances. We are now arriving at an observation that is central for the method we are developing. Every term in Eq. (21) is either independent of M, or proportional to M^{-1}. The covariance is thus of the form

$$\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2);M\right] = \mathrm{A}(\mathbf{r}_1,\mathbf{r}_2) + \frac{1}{M}\mathrm{B}(\mathbf{r}_1,\mathbf{r}_2). \tag{22}$$

Suppose we know the covariance for two distinct random-catalog sizes $M = M_a$ and $M = M_b > M_a$. We readily see that Eq. (22) holds when

$$\begin{aligned} \mathrm{A}(\mathbf{r}_1,\mathbf{r}_2) &= \frac{M_b}{M_b-M_a}\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2);M_b\right] \\ &\quad - \frac{M_a}{M_b-M_a}\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2);M_a\right] \\ \mathrm{B}(\mathbf{r}_1,\mathbf{r}_2) &= \frac{M_aM_b}{M_b-M_a}\Big\{\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2);M_a\right] \\ &\quad - \mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1),\hat{\xi}(\mathbf{r}_2);M_b\right]\Big\}. \end{aligned} \tag{23}$$

To construct the covariance for an arbitrary value of M, it is sufficient to estimate the covariance for two smaller random-catalog sizes M_a and M_b. This is much cheaper than running the estimator with a large M. Equations (22) and (23) can then be used to construct the covariance for the actual random catalog size. This is the basic idea behind our proposed method.

2.3. Linear construction

We now consider how M_a and M_b should be chosen. The largest reduction in computational cost is obtained with $M_a = 1$ and $M_b = 2$. With $M_b = M$ the method reduces to the conventional sample covariance.

We now argue in favor of selecting $M_b = 2M_a$. This allows us to use the random catalogs efficiently, as we now proceed to explain. For each mock data catalog we generate two independent random catalogs of same size M_a. We evaluate the $M = M_a$ covariance with either of the two sets, and take the average. This is the M_a covariance to enter the formula (23). For the $2M_a$ covariance, we take the combined data of the two M_a random catalogs. This procedure reduces the scatter of the estimate, compared to generating a new $2M_a$ random catalog, since the correlated fluctuations cancel. We return to this point in Sect. 3 where we consider the accuracy of the estimated covariance quantitatively. To further save CPU time we save the DR and RR pair counts from the $M = M_a$ simulations, and construct the $M = 2M_a$ 2PCF from the saved pair counts, saving the CPU cost of another run with $M = 2M_a$. Throughout the rest of this paper we set $M_b = 2M_a$.

Our method is summarized formally as follows. We denote by $\hat{\mathrm{C}}_{ij}$ an estimated covariance matrix between two elements $\xi(\mathbf{r}_i), \xi(\mathbf{r}_j)$ of the correlation function. Indices i, j may refer to different distance bins, or to elements picked from different multipoles.

We denote the correlation function estimated from a data catalog D and a random catalog R as $\hat{\xi}(\mathrm{D},\mathrm{R})$, and the sample covariance over the set of mocks as $\hat{C}_{ij}(\hat{\xi})$. We now have one data catalog D and two random catalogs R_1 and R_2, both of size M_a. We construct estimates for the covariances with $M = M_a$ and $M = 2M_a$ (which we denote by $\hat{\mathrm{C}}^a_{ij}$ and $\hat{\mathrm{C}}^b_{ij}$, respectively) as

$$\begin{aligned} \hat{\mathrm{C}}^a_{ij} &:= \tfrac{1}{2}\hat{C}_{ij}\left[\hat{\xi}(\mathrm{D},\mathrm{R}_1)\right] + \tfrac{1}{2}\hat{C}_{ij}\left[\hat{\xi}(\mathrm{D},\mathrm{R}_2)\right] \\ \hat{\mathrm{C}}^b_{ij} &:= \hat{C}_{ij}\left[\hat{\xi}(\mathrm{D},\mathrm{R}_1\cup\mathrm{R}_2)\right]. \end{aligned} \tag{24}$$

From these we construct two component matrices

$$\hat{\mathrm{A}}_{ij} := 2\hat{\mathrm{C}}^{b}_{ij} - \hat{\mathrm{C}}^{a}_{ij}$$
$$\hat{\mathrm{B}}_{ij} := 2M_a \left[\hat{\mathrm{C}}^{a}_{ij} - \hat{\mathrm{C}}^{b}_{ij}\right] \quad (25)$$

and the final covariance as

$$\hat{\mathrm{C}}^{\mathrm{LC}}_{ij} := \hat{\mathrm{A}}_{ij} + M^{-1}\hat{\mathrm{B}}_{ij}. \quad (26)$$

We refer to the covariance of Eq. (26) as the linear construction (LC) covariance.

The computational complexity of the Landy–Szalay estimator is roughly proportional to $\frac{1}{2}N_{\mathrm{d}}^2(1+3M)$, of which $\frac{1}{2}N_{\mathrm{d}}^2$ goes to DD pairs, $N_{\mathrm{d}}^2 M$ to DR pairs, and $\frac{1}{2}N_{\mathrm{d}}^2 M$ to RR pairs. The cost of our proposed approach is $2\cdot\frac{1}{2}N_{\mathrm{d}}^2(1+3M_a)$ per realization. This is assuming the 2PCF estimation code is run twice, which involves counting the DD pairs twice. If that is avoided, the cost is further reduced to $\frac{1}{2}N_{\mathrm{d}}^2(1+6M_a)$. With $M = 50$ and $M_a = 1$, the gain with respect to sample covariance is a factor of $151/7 \approx 21.6$, and increases with increasing M. Moreover, our result readily yields an extrapolation to infinite M, that is, we know what would be the estimator variance if we could use an infinite number of random points, and how much the variance with a finite M differs from that.

It is important to note that the presented derivation is based on very general assumptions on how the Landy–Szalay estimator is build from pair counts, and on the definition of variance. We do not make assumptions on the survey geometry, or which physical processes cause the scatter in the random counts, nor do we assume Gaussianity. The proposed method is thus valid for a very wide class of galaxy distributions.

Another important aspect to note is that the same procedure can be applied to any linear function of the correlation function. The only requirement is that the decomposition of Eq. (22) remains valid. In particular, the method applies as such to the multipoles $\xi_\ell(\mathbf{r})$ of the correlation function, and to the projected correlation function, since both are linear functions of the underlying 2-dimensional correlation function. It also applies to a rebinned correlation function.

3. Error analysis

We note that $\hat{\mathrm{C}}^{\mathrm{LC}}$ is a noisy estimate of the underlying true covariance C. Thus it is itself a random variable, and we can define a covariance for it. In the following we analyze the error of the covariance estimate, and derive a covariance of covariance for both the sample covariance and the LC covariance.

3.1. Gaussian distribution

We consider first the general case of four random variables x, y, w, z. For each of these we assume to have N independent realizations. An unbiased estimate of the covariance $\mathrm{C}(x, y)$ is obtained as

$$\hat{C}(x,y) := \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}), \quad (27)$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, and similarly for $\bar{y}$. It can be shown that

$$\mathrm{cov}\left[\hat{C}(x,y),\hat{C}(z,w)\right] = \frac{1}{N}\langle x'y'z'w'\rangle - \frac{1}{N}\langle x'y'\rangle\langle z'w'\rangle + \frac{1}{N(N-1)}\left(\langle x'z'\rangle\langle y'w'\rangle + \langle x'w'\rangle\langle y'z'\rangle\right), \quad (28)$$

Table 1. Symbols used in this work.

Symbol	Meaning
N_d	Number of objects in the data catalog
N_r	Number of objects in the random catalog
M	Ratio $M = N_r/N_d$, split factor
M_a	Reduced random catalog size
N	Number of realizations
$\xi(\mathbf{r})$	True two-point correlation function
$\hat{\xi}(\mathbf{r})$	Estimate of the correlation function
$\mathbf{r}$	Generalized (1D,2D, 3D) bin
C	True covariance
$\hat{\mathrm{C}}$	Estimated covariance
$\hat{C}(x,y)$	Numerical covariance of x, y
$\hat{\mathrm{C}}_{\mathrm{ref}}$	Reference covariance
$\tilde{\mathrm{C}}$	Covariance normalized to diagonal=1
A, B	Component matrices of the LC method
$\hat{\mathrm{A}}, \hat{\mathrm{B}}$	Estimates of the component matrices
$\hat{\mathrm{C}}^{a}_{ij}, \hat{\mathrm{C}}^{b}_{ij}$	Covariance estimates for $M = M_a$ and $M = 2M_a$
D, $\hat{\mathrm{D}}$	Combination $\mathrm{A} + \mathrm{B}/(2M_a)$ and its estimate
$\hat{\mathrm{C}}^{\mathrm{Smp}}$	Sample covariance
$\hat{\mathrm{C}}^{\mathrm{LC}}$	LC covariance
β	Linearized data model
F	$\beta^T\mathrm{C}^{-1}\beta$, inverse parameter covariance

where x', y', z', w' represent a deviation from the distribution mean, $x' = x - \langle x\rangle$. This is a general result that does not assume Gaussianity. For a Gaussian distribution

$$\langle x'y'z'w'\rangle = \langle x'y'\rangle\langle z'w'\rangle + \langle x'z'\rangle\langle y'w'\rangle + \langle x'w'\rangle\langle y'z'\rangle. \quad (29)$$

If x, y, z, w are Gaussian distributed, Eq. (28) simplifies into

$$\mathrm{cov}\left[\hat{C}(x,y),\hat{C}(z,w)\right] = \frac{1}{N-1}\left[\mathrm{C}(x,z)\mathrm{C}(y,w) + \mathrm{C}(x,w)\mathrm{C}(y,z)\right]. \quad (30)$$

3.2. Sample covariance

We can readily apply the results from above to the sample covariance. We take x, y, z, w to represent elements of the correlation function, as estimated through Landy–Szalay. We denote these elements by $\hat{\xi}_i, \hat{\xi}_j, \hat{\xi}_k, \hat{\xi}_l$. Different indices refer both to different distance bins, and to elements picked from different multipoles.

We denote the sample covariance for brevity by $\hat{\mathrm{C}}^{\mathrm{Smp}}_{ij} \equiv \hat{C}(\hat{\xi}_i, \hat{\xi}_j)$. Assuming that the $\hat{\xi}$ estimates follow the Gaussian distribution, the covariance of the sample covariance is

$$\mathrm{cov}\left(\hat{\mathrm{C}}^{\mathrm{Smp}}_{ij},\hat{\mathrm{C}}^{\mathrm{Smp}}_{kl}\right) = \frac{1}{N-1}(\mathrm{C}_{ik}\mathrm{C}_{jl} + \mathrm{C}_{il}\mathrm{C}_{jk}) \approx \frac{1}{N-1}\left(\hat{\mathrm{C}}^{\mathrm{Smp}}_{ik}\hat{\mathrm{C}}^{\mathrm{Smp}}_{jl} + \hat{\mathrm{C}}^{\mathrm{Smp}}_{il}\hat{\mathrm{C}}^{\mathrm{Smp}}_{jk}\right). \quad (31)$$

In particular for the diagonal elements

$$\mathrm{cov}\left(\hat{\mathrm{C}}^{\mathrm{Smp}}_{ii},\hat{\mathrm{C}}^{\mathrm{Smp}}_{kk}\right) = \frac{2}{N-1}\left[\mathrm{C}(\xi_i,\xi_k)\right]^2 \approx \frac{2\left(\hat{\mathrm{C}}^{\mathrm{Smp}}_{ik}\right)^2}{N-1}. \quad (32)$$

The $1\,\sigma$ uncertainty of a diagonal element of the sample covariance matrix is a fraction of $\sqrt{2/(N-1)}$ of the diagonal element itself. For $N = 5000$ this gives a 2% error ($1\,\sigma$). The off-diagonal part inherits the correlated structure of the covariance.

3.3. Linear construction

We can make use of Eq. (30) to derive covariance of covariance for the LC method as well. The basic data sets are now two sets of $M = M_a$ estimates of the correlation function, which we denote by x and x'. Again we assume that x, x' follow a Gaussian distribution. As described in Sect. 2.2, the $M = M_a$ covariance is constructed as

$$\hat{\mathrm{C}}^a_{ij} = \tfrac{1}{2}\left[\hat{C}(x_i, x_j) + \hat{C}(x'_i, x'_j)\right]. \tag{33}$$

For $M = 2M_a$ we construct the correlation function from the combined pair counts of the $M = M_a$ case. For large pair counts (when $\alpha, \beta, \gamma \ll 1$) $\hat{\xi}_i \approx \frac{1}{2}(x_i + x'_i)$, as we see from Eq. (12). The covariance is then

$$\begin{aligned}\hat{\mathrm{C}}^b_{ij} &= \hat{C}\left[\tfrac{1}{2}(x_i + x'_i), \tfrac{1}{2}(x_j + x'_j)\right] \\ &= \frac{1}{4}\left[\hat{C}(x_i, x_j) + \hat{C}(x'_i, x'_j) + \hat{C}(x_i, x'_j) + \hat{C}(x'_i, x_j)\right].\end{aligned} \tag{34}$$

The LC covariance for arbitrary M is constructed as

$$\hat{\mathrm{C}}^{\mathrm{LC}}_{ij} = \hat{\mathrm{A}}_{ij} + M^{-1}\hat{\mathrm{B}}_{ij}, \tag{35}$$

where now

$$\begin{aligned}\hat{\mathrm{A}}_{ij} &= 2\hat{\mathrm{C}}^b_{ij} - \hat{\mathrm{C}}^a_{ij} = \tfrac{1}{2}\hat{C}(x_i, x'_j) + \tfrac{1}{2}\hat{C}(x'_i, x_j) \\ \hat{\mathrm{B}}_{ij} &= 2M_a(\hat{\mathrm{C}}^a_{ij} - \hat{\mathrm{C}}^b_{ij}) \\ &= \tfrac{1}{2}M_a\left[\hat{C}(x_i, x_j) + \hat{C}(x'_i, x'_j) - \hat{C}(x_i, x'_j) - \hat{C}(x'_i, x_j)\right].\end{aligned} \tag{36}$$

Here we see the importance of constructing $\hat{\mathrm{C}}^a_{ij}$ and $\hat{\mathrm{C}}^b_{ij}$ from the same pair counts: The auto-correlation terms in $\hat{\mathrm{A}}_{ij}$ cancel out. In terms of x, x' the LC covariance is then

$$\begin{aligned}\hat{\mathrm{C}}^{\mathrm{LC}}_{ij} &= \tfrac{1}{2}\left(1 - \frac{M_a}{M}\right)\left[\hat{C}(x_i, x'_j) + \hat{C}(x'_i, x_j)\right] \\ &\quad + \frac{M_a}{2M}\left[\hat{C}(x_i, x_j) + \hat{C}(x'_i, x'_j)\right].\end{aligned} \tag{37}$$

We are now ready to construct the covariance of the LC covariance. Correlating the four terms of Eq. (37) yields 16 terms, each of which, with the use of Eq. (30), splits further into two terms. Taking into account that x and x' have identical statistical properties we finally arrive at

$$\begin{aligned}&\mathrm{cov}\left(\hat{\mathrm{C}}^{\mathrm{LC}}_{ij}, \hat{\mathrm{C}}^{\mathrm{LC}}_{kl}\right) \\ &= \frac{1}{N-1}\left[\mathrm{D}_{ik}\mathrm{D}_{jl} + \mathrm{D}_{il}\mathrm{D}_{jk} + \left(\frac{1}{2M_a} - \frac{1}{M}\right)^2(\mathrm{B}_{ik}\mathrm{B}_{jl} + \mathrm{B}_{il}\mathrm{B}_{kl})\right],\end{aligned} \tag{38}$$

where we have defined

$$\mathrm{D}_{ik} := \mathrm{A}_{ik} + \frac{1}{2M_a}\mathrm{B}_{ik}, \tag{39}$$

and A, B are ensemble average versions of (36). Using $\mathrm{C}_{ij} = \mathrm{A}_{ij} + M^{-1}\mathrm{B}_{ij}$ the result can be worked into the alternative form

$$\begin{aligned}&\mathrm{cov}\left(\hat{\mathrm{C}}^{\mathrm{LC}}_{ij}, \hat{\mathrm{C}}^{\mathrm{LC}}_{kl}\right) \\ &= \frac{1}{N-1}\Big[\mathrm{C}_{ik}\mathrm{C}_{jl} + \mathrm{C}_{il}\mathrm{C}_{jk} \\ &\quad + \left(\frac{1}{2M_a} - \frac{1}{M}\right)(\mathrm{D}_{ik}\mathrm{B}_{jl} + \mathrm{B}_{ik}\mathrm{D}_{jl} + \mathrm{D}_{il}\mathrm{B}_{jk} + \mathrm{B}_{il}\mathrm{D}_{jk})\Big].\end{aligned} \tag{40}$$

This allows a direct comparison with the sample covariance (Eq. (31)). The first line equals the covariance of the sample covariance. The second line represents additional error due to the reduced random catalog. When $M = 2M_a$, the LC covariance becomes equivalent to the sample covariance.

For all of the elements of Eqs. (40) or (38) we already have an estimate: $\mathrm{C}_{ik} \approx \hat{\mathrm{C}}^{\mathrm{LC}}_{ik}$, $\mathrm{B}_{ik} \approx \hat{\mathrm{B}}_{ik}$, $\mathrm{D}_{ik} \approx \hat{\mathrm{A}}_{ik} + \hat{\mathrm{B}}_{ik}/(2M_a)$. Thus we have a practical way of estimating the error of the LC covariance estimate.

3.4. Precision matrix and parameter estimation

The covariance matrix provides an account of the uncertainty in the 2PCF estimate. In many applications one is more interested in the inverse covariance, or the precision matrix

$$\Psi := C^{-1}. \tag{41}$$

The precision matrix enters a likelihood model, and is an ingredient in a maximum-likelihood parameter estimate. The properties of the precision matrix, when computed from the sample covariance, are relatively well understood. The inverse sample covariance is biased, but the bias can be corrected for with a multiplicative correction factor that only depends on the length of the data vector and on the available number of samples (see Anderson 2003; Hartlap et al. 2007, and references therein).

The effect of the accuracy of the precision matrix on parameter estimation has been studied by Taylor et al. (2013), Taylor & Joachimi (2014), Dodelson & Schneider (2013), Percival et al. (2014, 2022) and Sellentin & Heavens (2016). Taylor et al. (2013) present a remarkably simple result for the variance of the trace of the precision matrix. Dodelson & Schneider (2013) and Percival et al. (2014) compute the expected increase of estimated parameter errorbars due to the propagation of the sampling error of the covariance matrix. The increase is captured in a multiplicative factor that depends on the length of the data vector and on the number of independent parameters. Sellentin & Heavens (2016) use a fully Bayesian approach to incorporate the uncertainty of the estimated covariance into the likelihood function, for a more realistic likelihood which takes the form of a t-distribution. To have a clear interpretation of parameter posteriors in the case of a sample covariance matrix, Percival et al. (2022) propose a formulation of Bayesian priors that makes the parameter posteriors to match those in a frequentist approach of Dodelson & Schneider (2013) or Percival et al. (2014).

None of these results, unfortunately, generalize for the LC covariance, without further assumptions on the survey characteristics or on the parametric model. However, we do have the covariance of covariance, which can be used to assess the impact of covariance accuracy to a specific application, once the details are known. In the following we present some general observations.

One important aspect to note is that the LC covariance cannot be guaranteed to be positive-definite under all circumstances. This follows from the fact that the component matrix $\hat{\mathrm{A}}$ is constructed as a difference between two numerical covariances. If the actual covariance matrix is close to singular, random fluctuations in the numerical estimate may bring the smallest eigenvalues on the negative side. We recommend that if the inverse covariance is needed, the eigenspectrum of the matrix is verified first.

The precision matrix can be expanded as Taylor series as

$$\hat{\mathrm{C}}^{-1} = (\mathrm{C} + \Delta)^{-1} \approx \mathrm{C}^{-1} - \mathrm{C}^{-1}\Delta\,\mathrm{C}^{-1} + \mathrm{C}^{-1}\Delta\,\mathrm{C}^{-1}\Delta\,\mathrm{C}^{-1}, \tag{42}$$

where C is the true covariance and Δ the deviation of the estimate from it. The last term is the source of bias in the precision

matrix, which exists even if the covariance estimate is unbiased ($\langle\Delta\rangle = 0$). A bias in the precision matrix does not, however, translate into a bias in parameter estimation. The maximum-likelihood parameter estimate (without prior) is given by

$$\hat{p} = (\beta^T \hat{\mathrm{C}}^{-1}\beta)^{-1}\beta^T \hat{\mathrm{C}}^{-1} y, \tag{43}$$

where $\hat{p}$ (length n_p) represents the vector or estimated parameters, y (length n_b) is the data vector, $\hat{\mathrm{C}}$ is the covariance estimate, and

$$\beta_{i\alpha} = \frac{\partial y_i}{\partial p_\alpha} \tag{44}$$

is the linearized data model connecting the parameters to the data. One readily sees that the parameter estimate of Eq. (43) is unbiased regardless of $\hat{\mathrm{C}}$, and, if the covariance is biased by a multiplicative factor, the estimate is actually unaffected. It is therefore more interesting to look at the parameter covariance than at the bias of the precision matrix alone.

Following the example of Hartlap et al. (2007) we now insert the expansion of Eq. (42) into the parameter estimate of Eq. (43). We obtain for the parameter covariance

$$\begin{aligned}\langle \delta p \delta p^T\rangle = {} & \mathsf{F}^{-1} + \mathsf{F}^{-1}\beta^T \mathrm{C}^{-1}\langle \Delta\, \mathrm{C}^{-1}\Delta\rangle \mathrm{C}^{-1}\beta \mathsf{F}^{-1} \\ & - \mathsf{F}^{-1}\beta^T \mathrm{C}^{-1}\langle \Delta\, \mathrm{C}^{-1}\beta \mathsf{F}^{-1}\beta^T \mathrm{C}^{-1}\Delta\rangle \mathrm{C}^{-1}\beta F^{-1},\end{aligned} \tag{45}$$

where

$$\mathsf{F} := \beta^T \mathrm{C}^{-1}\beta. \tag{46}$$

If the covariance of covariance is of the general form

$$\langle \Delta_{ij}\Delta_{kl}\rangle = \frac{1}{N-1}(\mathrm{U}_{ik}\mathrm{V}_{jl} + \mathrm{U}_{il}\mathrm{V}_{jk}) \tag{47}$$

(as is the case for both sample covariance and LC) where U and V are arbitrary matrices, we find

$$\langle \Delta \mathrm{X} \Delta\rangle_{ij} = \frac{1}{N-1}[(\mathrm{U}\mathrm{X}^T\mathrm{V})_{ij} + \mathrm{U}_{ij}\mathrm{Tr}(\mathrm{X}^T\mathrm{V})], \tag{48}$$

where again X is an arbitrary matrix. We use Eq. (40) in combination with Eqs. (45) and (48) to derive for the parameter covariance the result

$$\begin{aligned}\langle \delta p \delta p^T\rangle = {} & \mathsf{F}^{-1}\left(1 + \frac{n_\mathrm{d} - n_\mathrm{p}}{N-1}\right) \\ & + \left(\tfrac{1}{2} - \tfrac{M_a}{M}\right)\frac{1}{N-1}\Big\{\mathsf{F}^{-1}\mathsf{R}\mathsf{F}^{-1} + \mathsf{F}^{-1}\mathsf{R}^T\mathsf{F}^{-1} \\ & - \mathsf{F}^{-1}\mathsf{P}\mathsf{F}^{-1}\mathsf{Q}\mathsf{F}^{-1} - \mathsf{F}^{-1}\mathsf{Q}\mathsf{F}^{-1}\mathsf{P}\mathsf{F}^{-1} \\ & + \mathsf{F}^{-1}\mathsf{P}\mathsf{F}^{-1}\left[\mathrm{Tr}(\mathrm{C}^{-1}\mathrm{D}) - \mathrm{Tr}(\mathsf{F}^{-1}\mathsf{Q})\right] \\ & + \mathsf{F}^{-1}\mathsf{Q}\mathsf{F}^{-1}\left[\mathrm{Tr}(\mathrm{C}^{-1}\mathrm{B}) - \mathrm{Tr}(\mathsf{F}^{-1}\mathsf{P})\right]\Big\},\end{aligned} \tag{49}$$

where

$$\begin{aligned}\mathsf{P} &:= \beta^T \mathrm{C}^{-1}\mathrm{B}\mathrm{C}^{-1}\beta, \\ \mathsf{Q} &:= \beta^T \mathrm{C}^{-1}\mathrm{D}\mathrm{C}^{-1}\beta, \\ \mathsf{R} &:= \beta^T \mathrm{C}^{-1}\mathrm{B}\mathrm{C}^{-1}\mathrm{D}\mathrm{C}^{-1}\beta.\end{aligned} \tag{50}$$

F^{-1} is the parameter covariance in the case where the data covariance C is known exactly. The first term represents the parameter covariance for sample covariance, a result in line with Dodelson & Schneider (2013). The rest is additional scatter specific for the LC method, and is dependent on the parametric model. Again we see that the additional terms vanish with $M = 2M_a$. Once the parametric model and β are fixed, and one has an estimate for C in the form of the LC covariance, Eq. (49) provides a practical recipe for estimating the parameter covariance.

4. Simulations

4.1. Cosmological mocks

To validate the LC method, we applied it to the computation of the 2PCF covariance matrix of simulated dark matter halo catalogs, and compared it to their sample covariance. We used mock catalogs produced with version 4.3 of PINOCCHIO[1] (PINpointing Orbit Crossing Collapsed HIerarchical Objects) algorithm (Monaco et al. 2002; Munari et al. 2017). This code is based on Lagrangian perturbation theory, ellipsoidal collapse and excursion sets approach. It is able to generate catalogs of dark matter halos, both in periodic boxes and in the past light cone, that closely match the mass function and the clustering of simulated halos without running a full N-body simulation. The particular configuration we used (see Colavincenzo et al. 2019) was run with ΛCDM cosmology using parameter values presented in Table 2. The simulation box had sides of length $L = 1500\,h^{-1}$ Mpc sampled with 1000^3 particles of mass $2.67 \times 10^{11}\,h^{-1}\,M_\odot$. The smallest identified halos consisted of 30 particles, which translates to masses of $8.01 \times 10^{12}\,h^{-1}\,M_\odot$. The mock catalogs we used correspond to a snapshot of the simulation in a periodic box at redshift $z = 1$. The 10 000 realizations were run with the same configuration, but with different seeds for random numbers. As a consequence the number of halos in each PINOCCHIO realization is subject to sample variance. The mean number of halos in a box is 780 789 and varies from box to box by $^{+0.3}_{-0.4}$%. This corresponds to a number density of $2.3 \times 10^{-4}\,(h^{-1}\,\mathrm{Mpc})^{-3}$.

The PINOCCHIO mocks contain the halo positions in real space and their peculiar velocities, in a periodic box. To imitate a real survey more closely we mapped the halo positions into redshift space. We worked within the plane-parallel assumption; we constructed a periodic redshift-space box by shifting the halo positions along the x-axis according to the peculiar velocity component along the same axis. In order to compute the correlation function multipoles, we must define the location of the observer with respect to the simulation box. To preserve the plane-parallel assumption, we moved the observation point along the x-axis to a distance of $10^6\,h^{-1}$ Mpc from the box.

To mimic the geometry of a tomographic survey with a limited redshift coverage we selected a slab-like subset of the full simulation box. The thickness of the slab is $L/5 = 300\,h^{-1}$ Mpc. This geometry is shown in Fig. 1. The mean number of halos in the slab is one fifth of that of the full box (156 158 objects) and varies by ±3%. For the corresponding random catalogs we generated random coordinates homogeneously inside the slab using the method `random.rand` of the `numpy` python library. The number of random points in each slab is M times the number of halos, so the size of each random catalog is also slightly different.

The area of the simulation slab corresponds to a solid angle of 1400 square degrees at $z = 1$, which is 9.4% of the 15 000 deg^2 sky coverage of the *Euclid* spectroscopic survey (Laureijs et al. 2011; Euclid Collaboration 2022). The thickness of the slab corresponds to a redshift bin of $\Delta z \approx 0.2$. The mean number of objects in the slab corresponds to 5% of the survey (30 million objects). The small number of objects in the simulation made it possible to construct the sample covariance for a large number of realizations, and thus to compare the accuracy and efficiency of the LC method against that of the sample covariance.

We had 10 000 halo catalog realizations at our disposal. We divide them into two sets of 5000 realizations. We computed

[1] https://github.com/pigimonaco/Pinocchio

Table 2. Parameter values for the PINOCCHIO simulation used in our analysis.

H_0	Ω_m	Ω_Λ	Ω_b	σ_8	n_s
0.695	0.285	0.715	0.044	0.828	0.9632

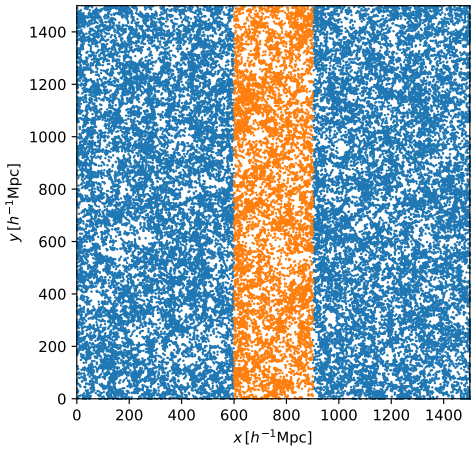

Fig. 1. Geometry of the mock catalogs used in our analysis. Blue points are the full simulation box and orange points are the slab we use for our analysis. Projected here is a slice of thickness of $100\,h^{-1}$ Mpc.

the sample covariance of one set, and used it as the reference covariance, against which we compared the other estimates. The reference covariance represents the best knowledge we have on the true covariance. We used the other set of 5000 realizations to compute both the LC covariance and the sample covariance, which we then compared with the reference covariance. This way we were able to estimate how much of the difference between the LC covariance and sample covariance was caused purely by the limited number of realizations.

We generated 10 000 random catalogs of size $N_\mathrm{r} = 50N_\mathrm{d}$, 5000 for the reference covariance, and 5000 for the sample covariance LC was compared with. In addition we generated 10 000 random catalogs of size $N_\mathrm{r} = 1N_\mathrm{d}$. We used a set of 5000 to serve as catalog R_1 in Eq. (24), and another set of 5000 as catalog R_2.

4.2. Random mocks

In the case of PINOCCHIO mocks we do not know the actual covariance exactly. We can only compare against the reference covariance, which itself is estimated from a finite data set. To have a test case where we know the true covariance, we ran another simulation using a purely random distribution of points as our data catalog. For this purpose we generated another set of 10 000 random mocks and used these in the place of the data catalog. Otherwise the setup was exactly the same as with PINOCCHIO mocks. We used the same slab geometry and point density, with the exception that each data catalog (and correspondingly each random catalog) realization has the same number of points: $N_\mathrm{d} = 2.3 \times 10^{-4}\,(h^{-1}\,\mathrm{Mpc})^{-3} \times 6.75 \times 10^8\,(h^{-1}\,\mathrm{Mpc})^3 = 155\,250$. The correlation function for the random distribution is zero, and for the covariance, as well as for the covariance of the covariance, an analytic result can be derived. This allowed us to directly compare the estimated covariance against the expected result.

4.3. Constructing the covariance

To validate the LC method, we computed the correlation function of the mock catalogs and constructed the LC covariance. Since we were looking for maximal reduction in the computational cost, we set $M_a = 1$, i.e. we used random catalogs of same size as the data catalog.

We computed the correlation function of the simulated galaxy distribution using the 2PCF code developed for the *Euclid* mission. The code implements the Landy-Szalay estimator with split random catalog, and stores as a by-product the DD, DR, and RR pair counts, which we need for the construction of the LC covariance. We used the *Euclid* code to compute the 2-dimensional correlation function $\hat{\xi}(r, \mu)$, where r is the distance between a pair of galaxies, and μ is the cosine of the angle between the line-of-sight and the line segment connecting the galaxy pair. We used bin sizes $\Delta r = 1\,h^{-1}$ Mpc and $\Delta\mu = 0.01$, and computed the correlation function for the distance range $r \in [0, 200]\,h^{-1}$ Mpc. For some tests we needed also the 1-dimensional correlation function, which we obtained by coadding the pair counts in μ dimension. For each data catalog, we ran the code three times: once to construct the $M = 50$ correlation function, and twice with $M = 1$ random catalogs to produce the pair counts we needed for the construction of the LC covariance.

We constructed the LC and sample covariance estimates with an external code, which takes as input the precomputed pair counts. To ensure consistency, we recomputed the correlation functions from the pair counts. Having the precomputed pair counts on disk also left us the possibility of combining bins into wider ones. The run time of this external code is negligible, the CPU usage being fully dominated by the run-time of the *Euclid* code.

We computed the correlation function multipoles from the two-dimensional correlation function as

$$\xi_\ell(r) := \frac{2\ell + 1}{2} \int_{-1}^{1} \xi(r, \mu)\, P_\ell(\mu)\, \mathrm{d}\mu, \tag{51}$$

where $P_\ell(\mu)$ are Legendre polynomials ($\ell = 0, 2, 4$). The $M = 50$ correlation function multipoles, as estimated from the simulation slabs, are depicted in Fig. 2. We show the mean over 5000 realizations, and a single realization. For this small survey size, a single realization deviates strongly from the ensemble mean, and the errors are strongly correlated between distance bins.

The calculation of the covariance of covariance in Sect. 3 relies on the assumption that the elements of the correlation function follow a Gaussian distribution, at least approximatively. To verify the validity of this assumption, we plot the distributions of selected correlation function elements in Fig. 3. The assumption of approximative Gaussianity seems well justified. We note that Gaussianity is only required for the covariance of covariance to be valid. The LC covariance itself does not rely on any particular distribution.

5. Results

5.1. Random mocks

We begin by examining the one-dimensional 2PCF of the random mocks. As explained above, we ran tests using randomly distributed points in place of the data catalog. This has the benefit that we know exactly the expected correlation function (zero). We also have an accurate analytic estimate for the true

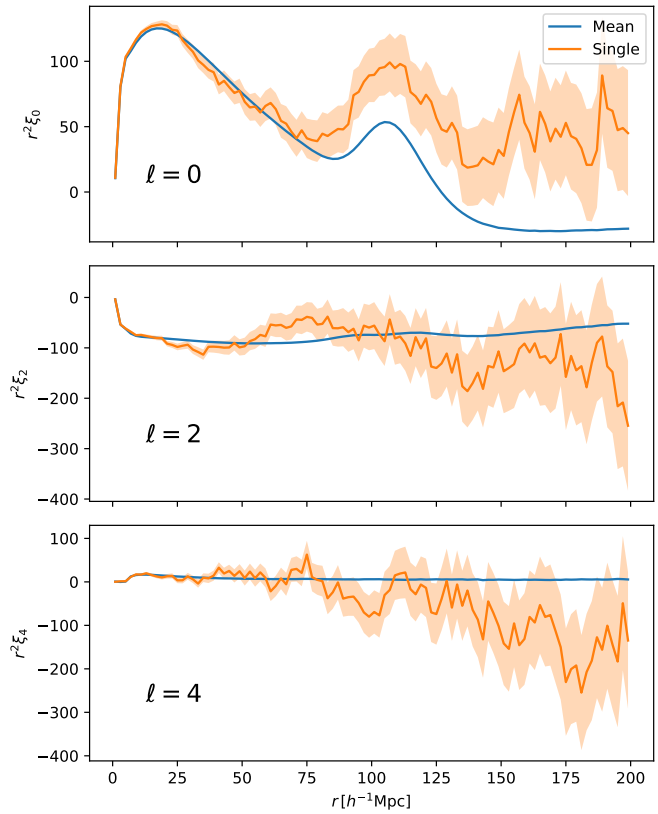

Fig. 2. Correlation function multipoles. Mean over 5000 PINOCCHIO realizations and a single realization. The shaded area around the single realization curve is the 1σ error envelope, computed as the standard deviation of the available realizations.

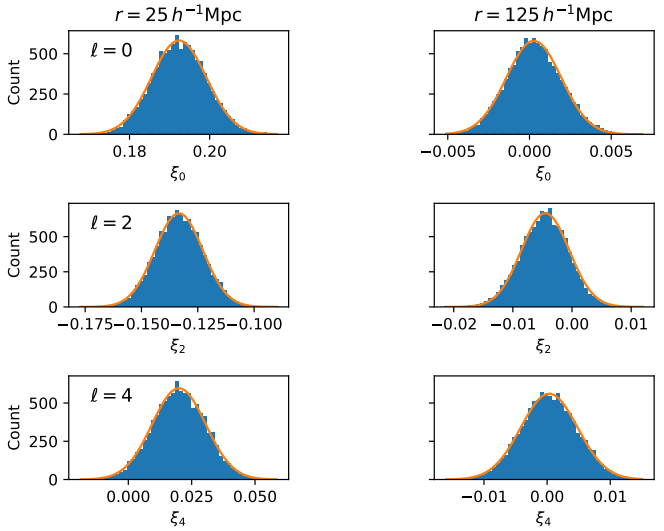

Fig. 3. Histogram of correlation function multipole values at $r = 25\,h^{-1}$ Mpc and $r = 125\,h^{-1}$ Mpc. Along with the histograms we show the corresponding best-fit Gaussian distribution in orange.

covariance. From Keihänen et al. (2019) we have

$$\begin{aligned}
&\mathrm{cov}\left[\hat{\xi}(\mathbf{r}_1), \hat{\xi}(\mathbf{r}_2)\right] \\
&= \frac{\delta_{12}}{G_\mathrm{p}(\mathbf{r}_2)}\left(\frac{2}{N_\mathrm{d}(N_\mathrm{d}-1)} + \frac{4}{N_\mathrm{d}N_\mathrm{r}} + \frac{2}{N_\mathrm{r}(N_\mathrm{d}-1)}\right) \\
&\approx \frac{\delta_{12}}{G_\mathrm{p}(\mathbf{r}_1)}\frac{2}{N_\mathrm{d}(N_\mathrm{d}-1)}\left(1 + 3M^{-1}\right),
\end{aligned} \tag{52}$$

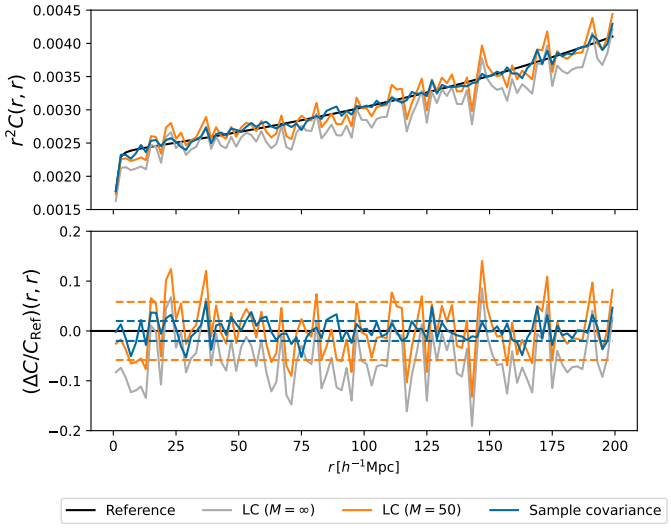

Fig. 4. Diagonal covariance from random mocks, one-dimensional case. *Top*: sample covariance, LC, $M = \infty$ limit, and theoretical prediction. *Bottom*: the relative errors, and theoretical 1σ error estimates.

where N_r is the number of random points and N_d the number of data points, and $G_\mathrm{p}(\mathbf{r})$ is the geometrical pair volume fraction defined as

$$G_\mathrm{p}(\mathbf{r}) = \left[\iint_V \mathrm{d}^3x_1\mathrm{d}^3x_2\right]^{-1} \iint_V W(x_1 - x_2 \in \mathbf{r})\mathrm{d}^3x_1\mathrm{d}^3x_2 \tag{53}$$

where the integrals cover the survey volume and $W(x_1 - x_2 \in \mathbf{r}) = 1$ if the pair distance falls in the distance bin of $\mathbf{r}$, and is zero otherwise. This allows us to directly compare the estimated covariance to the theoretical one.

Figure 4 shows the diagonal of the estimated covariance of the one-dimensional correlation function for $M = 50$, compared with the theoretical value of Eq. (52). We show also the $M \to \infty$ limit from the LC method. This represents the optimal covariance which we would have if we had an infinite random catalog. As expected, the $M = \infty$ curve lies slightly below the $M = 50$ curve. The difference is the additional uncertainty from the finite random catalog. Both the sample covariance and LC covariance agree very well with the expected covariance. It is also evident that the LC method results in larger scatter. The lower panel shows the relative difference with respect to the theoretical value, together with $1\,\sigma$ error bars derived from Eqs. (31) and (38). The error for the LC covariance, measured as the standard deviation, is 2.7 times that of the sample covariance, implying that more than 7 times more realizations are needed to reach the same level of accuracy. Fortunately, from the point of view of the LC method, this is an unrealistically pessimistic situation. This can be traced to the fact that correlations, which in a more realistic situation contribute significantly to the covariance, are nonexistent here. Thus the scatter of the random catalog, which in our method is large due to the small number of objects, contributes a large fraction of the total error. The situation looks very different when we move to realistic cosmological simulations with large correlations.

5.2. Cosmological mocks

As explained in Sect. 4.1, since we do not know the true covariance, we divided the available 10 000 realizations into two sets of 5000 realizations and used the sample covariance of the first

half as a reference. We constructed the covariance for $\ell = 0, 2, 4$ multipoles both through the sample covariance and with the LC method, with $M = 50$.

First we examine the convergence with respect to the number of realizations. This is shown in Fig. 5. We show the squared-sum difference with respect to the reference matrix, for the sample covariance and for LC. Because the bins at the smallest scales have only a few halos, we include the scales in the range 20–200 h^{-1} Mpc in the sum. We vary the number of realizations used for the covariance estimate under study, but the reference matrix in all cases is the same, based on the full set of 5000 realizations. All matrices have been normalized to the reference diagonal, in order to assign equal weights to all distance scales,

$$\hat{\mathrm{C}}_{ij}(\mathrm{normalized}) = \frac{\hat{\mathrm{C}}_{ij}}{\sqrt{\hat{\mathrm{C}}^{\mathrm{ref}}_{ii}\hat{\mathrm{C}}^{\mathrm{ref}}_{jj}}}. \tag{54}$$

We show the difference as a function of number of realizations, and as a function of CPU time. To further reduce the noise in the measurement we compute the convergence ten times and show the mean over these ten cases. Each case is obtained by randomly splitting the 10 000 PINOCCHIO realizations into two sets of 5000 realizations, one of which used to compute the reference covariance and the other one to compute the sample and LC covariances. The different splits overlap with each other, but even so the procedure significantly reduces the noise in the measured convergence. For the same number of realizations, the sample covariance gives a smaller uncertainty. One needs roughly 1.5 times the number of realizations with LC, to reach the same level of accuracy. In terms of CPU time spent, the situation is inverted. The LC covariance requires only 10% of the CPU cost of the sample covariance to reach the same accuracy.

Of the total wall-time of constructing the sample covariance, 90% is spent on counting the pairs. In the case of LC, this fraction is somewhat lower, 76%. Loading in the catalogs takes roughly the same fraction of time in both cases so the difference in efficiency seems to be in overheads such as code initialization. A possible optimization to reduce these overheads would be to compute all the thousands of 2PCF estimates during a single code execution instead of calling the code executable over and over again.

In Fig. 6 we show the diagonal of the covariance matrix monopole block, for the sample covariance and the LC estimate, along with the reference. We show also the $M = \infty$ limit of the LC covariance. In the lower panel we show the relative difference with respect to the reference covariance, and the theoretical prediction for the error, as given by Eqs. (31) and (38). Since we are looking at the difference with respect to the reference, the error level shown is the square root of the sum of the variances of the reference and the estimate in question.

Again, the LC estimate has more internal scatter than the sample covariance, but the difference between the methods is significantly smaller than in the case of random mocks. A more striking phenomenon is that the deviation from the reference is strongly correlated in distance, and the general trend of the deviation is very similar for the two estimation methods. In other words, the estimation error is dominated by a correlated error component that is independent of the chosen estimation method, when both estimates are constructed from the same data set. The common component dominates over the additional noise added by the LC method. The amplitude of the component is consistent with the predicted error level, indicating that it represents a random fluctuation.

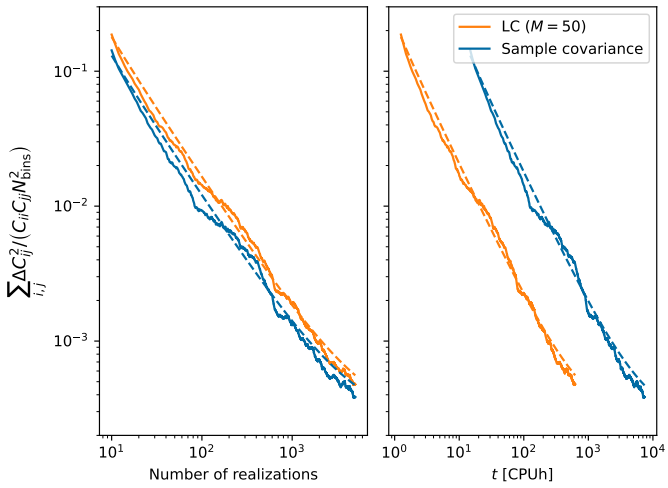

Fig. 5. Convergence of the covariance of the correlation function multipoles, with respect to the number of realizations (*left*) and CPU time (*right*). We use PINOCCHIO mocks and include scales of $r > 20\,h^{-1}$ Mpc. Dashed lines show the theoretical prediction.

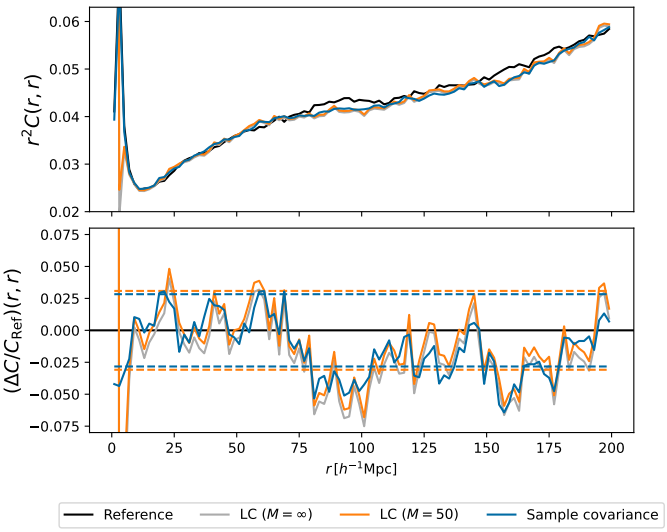

Fig. 6. Diagonal of the covariance for the correlation function monopole, PINOCCHIO mocks. On the top we show the LC and sample covariance estimates, and the $M = \infty$ limit. On the bottom we show the relative errors, and predicted 1σ error level.

Figure 7 shows the monopole block of the full LC covariance matrix as a two-dimensional plot. For plotting purposes we normalize the matrix by the diagonal of the reference covariance. There is significant off-diagonal component, showing that the error in the estimated Landy–Szalay correlation function is correlated from one distance bin to another, in line with Fig. 6. The middle panel shows the difference between the LC covariance and the reference. There is no obvious overall bias (which would show up as the over-representation of either the blue or the red color), but the region of correlated error is clearly visible around 100 h^{-1} Mpc. The bottom panel shows the difference between the LC and sample covariances from the same 5000 realizations. Here the structure is weaker, indicating that the correlated structure in the middle panel is for a large part common for the sample covariance and the LC estimate, as we already saw in Fig. 6.

We proceed to examine the structure of the LC covariance further. We show the $\hat{\mathrm{A}}$ and $\hat{\mathrm{B}}$ components for the full multipole covariance in Fig. 8, again normalized with the reference diagonal. The full covariance will be the combination $\hat{\mathrm{A}} + \hat{\mathrm{B}}/M$. We observe that the $\hat{\mathrm{B}}$ component is strongly diagonal-dominated, in contrast to the $\hat{\mathrm{A}}$ component, indicating that the finite random catalog mainly contributes uncorrelated noise to the 2PCF estimate, on top of the correlated error that arises from the data catalog. The unnormalized diagonals of all three multipole blocks, and their cross-components, are shown in Fig. 9.

The expectation value of the LC covariance in terms of pair-count covariances is given in Eq. (21). However, if we expand Eq. (26) (which defines the LC covariance) in terms of pair-count covariances, we find that the expansion includes more terms than Eq. (21). The expectation value of these additional terms vanishes, but when the covariance is estimated from a finite number of correlation function realizations, these terms differ from zero randomly. This raises the question whether leaving some or all of these zero-expectation-value terms out and constructing the covariance using the pair-count covariances directly would reduce noise in the covariance matrix estimate. We reconstructed the covariance matrix by including all the possible combinations of the zero-expectation-value terms, but it turned out that the most accurate combination is the one defined by Eq. (26). Even though the pair-count covariances do not affect the expectation value of the covariance matrix estimate, they do reduce its variance. This can be understood as follows: the zero-expectation terms are negatively correlated with some of the nonzero terms, and thus they help to cancel out part of the estimation noise.

5.3. Predictions from covariance of covariance

We now proceed to examine the accuracy of the LC covariance estimate in a more quantitative way. Here we make use of predictions of the theoretical covariance of covariance from Sect. 3.

We measure the accuracy of the covariance estimate, as the normalized sum-of-squares difference from the ensemble-average, over all covariance elements,

$$\chi^2_N := \frac{1}{N^2_{\mathrm{bin}}} \sum_{ij} \frac{1}{\hat{\mathrm{C}}^{\mathrm{ref}}_{ii} \hat{\mathrm{C}}^{\mathrm{ref}}_{jj}} (\hat{\mathrm{C}}_{ij} - \langle \mathrm{C}_{ij} \rangle)^2. \tag{55}$$

Here $\hat{\mathrm{C}}$ represents the covariance estimate, either LC or sample covariance, measured from N realizations (5000), and N_{bin} is the number of correlation function elements. In our baseline simulation $N_{\mathrm{bin}} = 540$ (3 multipoles and 180 distance bins). We normalized the sum by the diagonal of the reference covariance, to assign equal weights to all distance bins. Equation (55) expresses the accuracy of the covariance as a single number.

In terms of the covariance of covariance we have

$$\langle \chi^2_N \rangle = \frac{1}{N^2_{\mathrm{bin}}} \sum_{ij} \frac{\mathrm{cov}(\hat{\mathrm{C}}_{ij}, \hat{\mathrm{C}}_{ij})}{\hat{\mathrm{C}}^{\mathrm{ref}}_{ii} \hat{\mathrm{C}}^{\mathrm{ref}}_{jj}}. \tag{56}$$

Since the covariance of covariance scales as $1/(N-1)$, we can write this in terms of the $N = 2$ value as

$$\langle \chi^2_N \rangle = \frac{\langle \chi^2_2 \rangle}{N-1}. \tag{57}$$

For the sample covariance we now have

$$\langle \chi^2_2 \rangle^{\mathrm{Smp}} = \frac{1}{N^2_{\mathrm{bin}}} \sum_{ij} \left(\tilde{\mathrm{C}}_{ii} \tilde{\mathrm{C}}_{jj} + \tilde{\mathrm{C}}^2_{ij} \right), \tag{58}$$

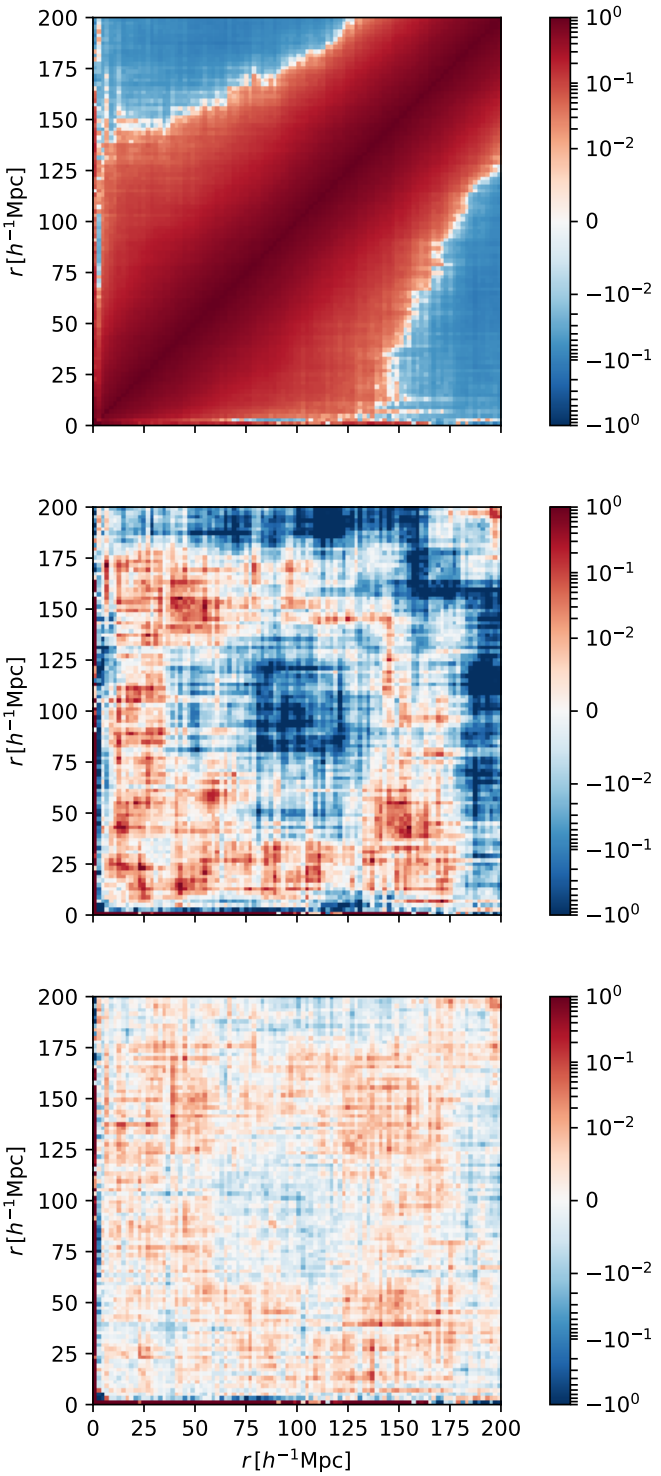

Fig. 7. Monopole block of the covariance matrix. *Top*: LC covariance matrix. *Middle*: difference between the LC covariance matrix and the reference. *Bottom*: difference between the LC and the sample covariance from same realizations. All are normalized by the diagonal elements of the reference matrix.

where we have absorbed the normalization into the covariance, and denoted the normalized covariance by $\tilde{\mathrm{C}}$. For LC we find

$$\langle \chi^2_2 \rangle^{\mathrm{LC}} = \frac{1}{N^2_{\mathrm{bin}}} \sum_{ij} \Big[\tilde{\mathrm{D}}_{ii} \tilde{\mathrm{D}}_{jj} + \tilde{\mathrm{D}}^2_{ij} + \left(\tfrac{1}{2} - M^{-1} \right)^2 \left(\tilde{\mathrm{B}}_{ii} \tilde{\mathrm{B}}_{jj} + \tilde{\mathrm{B}}^2_{ij} \right) \Big]. \tag{59}$$

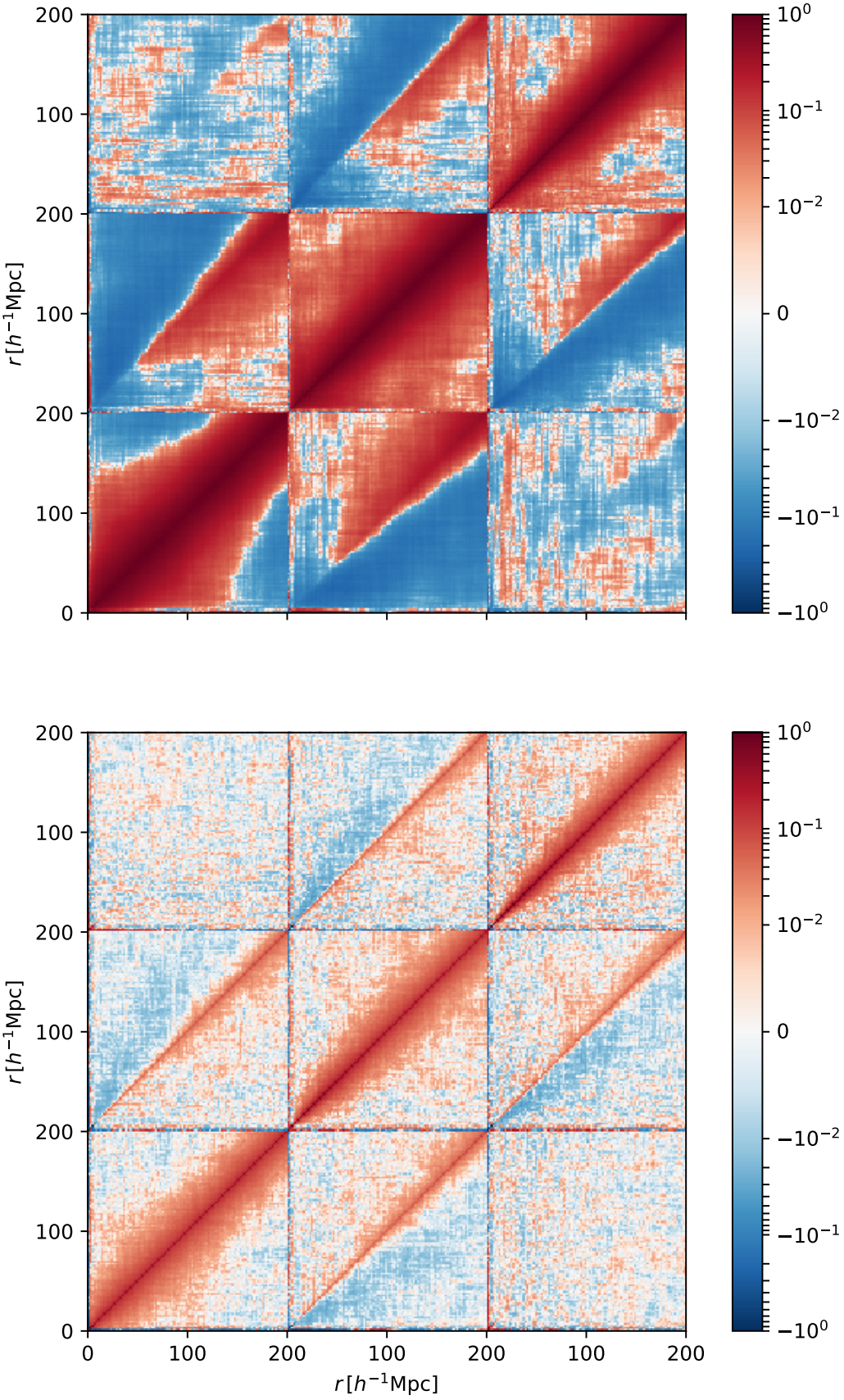

Fig. 8. Component matrices $\hat{\mathrm{A}}$ (*top*) and $\hat{\mathrm{B}}$ (*bottom*), for correlation function multipoles, measured from the PINOCCHIO mocks. The blocks from left to right and from the bottom to the top row correspond to $\ell = 0, 2, 4$ multipoles, respectively. Both are normalized by the diagonal elements of the reference matrix.

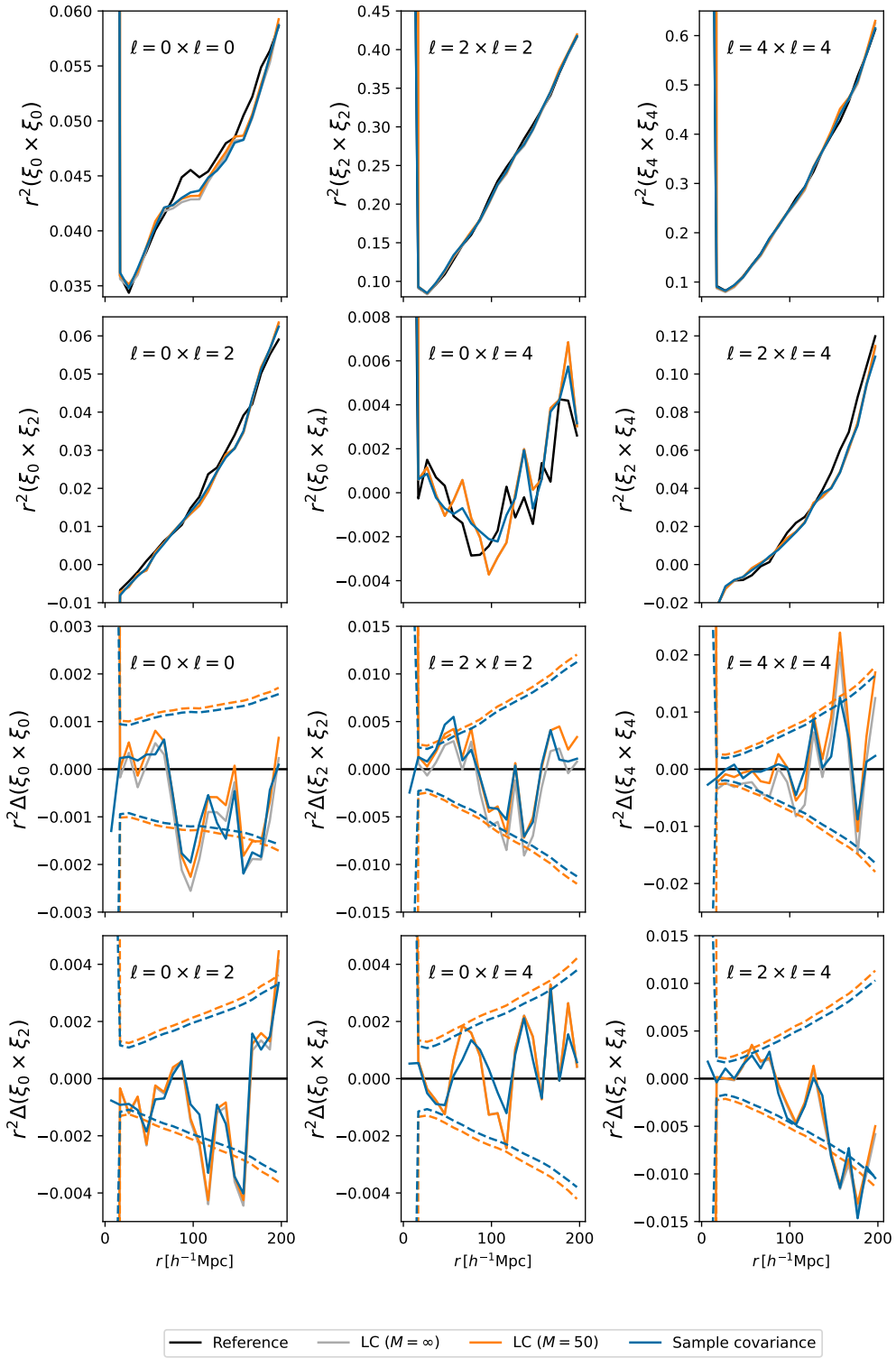

Fig. 9. Covariance diagonals for multipoles $\ell = 0, 2, 4$, and their cross-correlation, for PINOCCHIO. Sample covariance and LC. Two bottom rows show the difference between the reference and the estimate scaled by r^2. To reduce scatter in the curves all the quantities have been rebinned to bins of width of $10\,h^{-1}$ Mpc.

Here we have a practical way of predicting the estimation error for the LC and sample covariance, for different values of M. We can also easily predict the effect of rebinning the data into wider distance bins, simply by rebinning the covariance matrices and constructing the covariance of covariance from these.

In Table 3 we have collected statistics on the estimation methods, for a selected random catalog size (different values of M) and for different rebinning schemes. We have used $\hat{\mathrm{A}}$ and $\hat{\mathrm{B}}$ in the place of A and B, and $\hat{\mathrm{C}} \approx \hat{\mathrm{A}} + M^{-1}\hat{\mathrm{B}}$ in the place of C. We show the computational cost of pair counting in each of the cases, in the units of counting the pairs in one N_d data catalog. The cost of the LC method is the same in all cases, while the cost of the sample covariance scales as $1 + 3M$. The cost estimate ignores parts of the computation other than pair counting,

Table 3. Predictions for the variance of the covariance estimate, based on the covariance of covariance for PINOCCHIO mocks.

		Cost		χ^2_2		iFoM		
M	Δr	Smp	LC	Smp	LC	Smp	LC	Ratio
10	1	31	7	1.151	1.852	35.7	13.0	2.8
	2			1.124	1.582	34.8	11.1	3.1
	5			1.106	1.400	34.3	9.8	3.5
	10			1.102	1.328	34.2	9.3	3.7
	20			1.102	1.283	34.1	9.0	3.8
20	1	61	7	1.082	1.871	66.0	13.1	5.0
	2			1.076	1.591	65.6	11.1	5.9
	5			1.073	1.404	65.5	9.8	6.7
	10			1.076	1.331	65.7	9.3	7.0
	20			1.081	1.284	65.9	9.0	7.3
50	1	151	7	1.041	1.883	157	13.1	11.9
	2			1.048	1.597	158	11.1	14.1
	5			1.054	1.407	159	9.8	16.2
	10			1.061	1.332	160	9.3	17.2
	20			1.068	1.285	161	9.0	17.9
100	1	301	7	1.028	1.887	309	13.2	23.4
	2			1.038	1.599	313	11.2	27.9
	5			1.048	1.408	315	9.9	32.0
	10			1.056	1.333	318	9.3	34.1
	20			1.064	1.286	320	9.0	35.6

Notes. Sample covariance (Smp) and LC covariance with M_a=1 are compared. The columns are: size of random catalog, parametrized as $M = N_r/N_d$; bin size Δr in units of h^{-1} Mpc; computational cost of pair counting per realization, in units of the pair count cost of the data catalog; χ^2_2, variance of the covariance estimate per bin for $N = 2$ realizations and for distance scales 20–200 h^{-1} Mpc; inverse figure-of-merit, product of χ^2_2 and computational cost; ratio of sample-covariance iFoM to that of the LC.

for instance disk I/O and various overheads, thus exaggerating the difference between the methods. The estimator variance is expressed as χ^2_2. The standard deviation of a covariance estimate is obtained from this as $\sqrt{\chi^2_2/(N-1)}$. We show also an inverse figure-of-merit (iFoM) constructed as the product of the pair-count cost and the χ^2_2 value. Since the estimator variance decreases proportionally to the inverse of the number of realizations N, while the computation time grows proportional to it, this is an N-independent measure of the estimator efficiency. A smaller value indicates a more efficient estimation. The value of iFoM can be interpreted as the computational cost of reaching $\chi^2_2 = 1$. The last column shows the ratio of the sample-covariance iFoM to that of the LC covariance, and is measure of the gain from the LC method.

The relative efficiency of the LC method increases with increasing M, as the computational cost of the sample covariance becomes larger. We observe also that for given M, the LC covariance becomes more efficient in comparison to sample covariance, if we combine the distance bins into wider bins. With $M = 50$ and with 20 h^{-1} Mpc bins in distance, the efficiency ratio is 17.9, while with narrow 1 h^{-1} Mpc bins the ratio is 11.9.

The full covariance of covariance is a four-dimensional data object, and is difficult to visualize in its entirety. In the following we examine a two-dimensional subset. We focus on the diagonal of the monopole block of the covariance estimate (plotted in Fig. 6). This is a 1-dimensional data object, thus its covariance is a 2-dimensional matrix. In the following we refer to the covariance of the diagonal of the monopole block of a 2PCF covariance estimate as COVCOV for short. We plot the

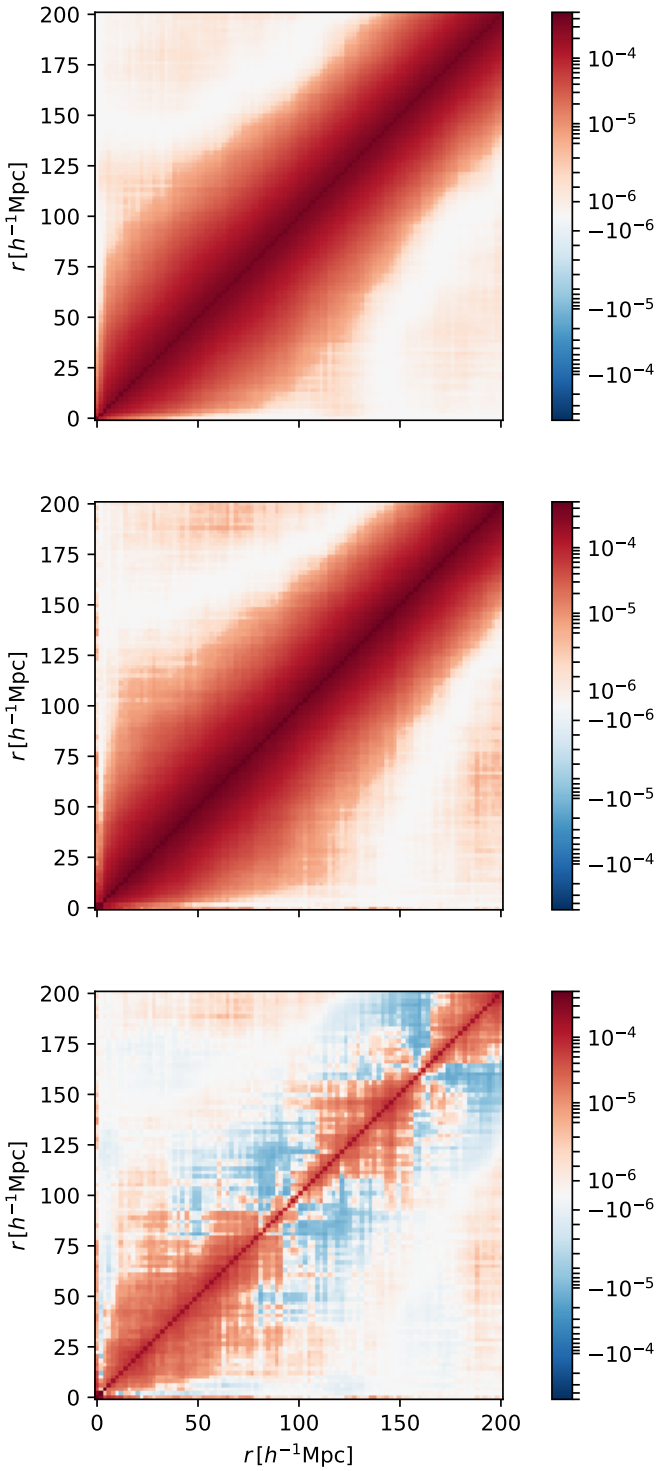

Fig. 10. Predicted covariance of the diagonal of the monopole block of the estimated 2PCF covariance (COVCOV). *Top*: COVCOV for sample covariance. *Middle*: COVCOV for LC. *Bottom*: difference of the two. From normalized covariances.

predicted COVCOV for the sample covariance and for LC in Fig. 10. In both cases, there is a significant off-diagonal structure, which is visually very similar between the two methods. This verifies our earlier observation that the estimation error is correlated between distance bins, and this correlation does not depend on the chosen method. Taking the difference between the COVCOV matrices, we see that the LC estimate has additional

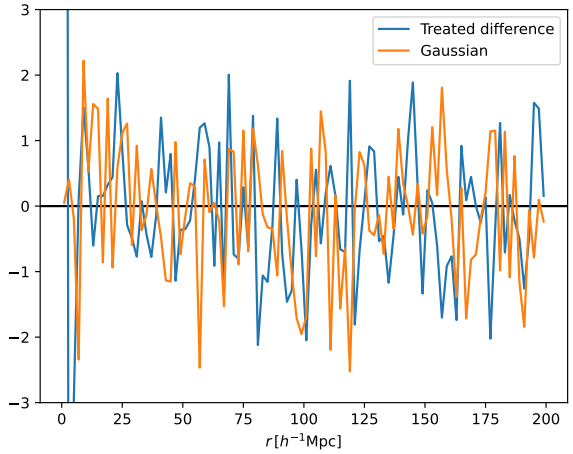

Fig. 11. Diagonal monopole difference treated with the inverse square root of COVCOV. For reference a Gaussian random vector of same length.

scatter compared to the sample covariance, but this additional error component is only weakly correlated from one distance bin to another.

As a final validation test we applied the inverse square root of the COVCOV to the difference between the LC covariance matrix diagonal and the corresponding reference (quantities plotted in Fig. 6). If the COVCOV correctly describes the errors in the estimated covariance, we expect to see an array of white noise with $\sigma = 1$. To account for the fact that the reference covariance has a covariance of its own, we took the COVCOV to be the sum of the reference and the LC COVCOV matrices. We computed the square root using the Schur method implemented in the `scipy` Python library. The resulting whitened data vector is shown in Fig. 11, along with a random realization of white noise. The data is visually indistinguishable from white noise, which is a valuable validation check. The similarity can also be confirmed by computing a normalized χ^2 value

$$\chi^2 = \frac{1}{N}\mathbf{v}^{\mathrm{T}}\mathbf{v}. \tag{60}$$

Here $\mathbf{v}$ is the data vector, N is the number of bins, and for a vector of Gaussian white noise we expect a value close to 1. We computed this value for scales $r > 20\,h^{-1}$ Mpc and obtained $\chi^2 = 0.95$ for the whitened data vector and $\chi^2 = 0.93$ for the Gaussian random vector.

6. Conclusions

We have presented a method for speeding up the computation of the galaxy 2PCF covariance. We have named the method as the Linear Construction, or LC method. We assume that the correlation function is estimated through Landy–Szalay estimator, with a split random catalog. The proposed method applies both to the raw (1- or 2-dimensional) correlation function and to its multipoles.

The proposed method provides an unbiased estimate of the covariance for a split random catalog, that is, for a case where the random-random pair count is constructed as the coadded sum of many small subcatalogs. Since we know that the splitting only weakly affects the 2PCF estimation error, we expect that the LC covariance provides a good approximation also when the RR pairs are counted from the full catalog. This can be traced to the fact that, for large random catalogs, the 2PCF estimation error is dominated by the variance of the galaxy sample, and the secondary error term is related to the data-random pair count, both of which are unaffected by splitting. The scatter of the random-random count plays a minor role.

The computational cost of the LC method per realization, for a random catalog M times the size of the data catalog, is a factor of $(1 + 3M)/7$ lower than that of the sample covariance. For $M = 50$ this yields a factor of 21.6 speedup. However, a larger number of realizations is needed, to compensate for the increased scatter in the estimate. In our simulations, 1.2–1.8 times higher a number of realizations was needed to reach a given level of accuracy, depending on bin size. This taken into account, the net cost reduction for $M = 50$ is a factor of 11.9–17.9. The efficiency increases with increasing bin width. In practice, we observe a halved speedup due to the heavy overhead associated with the handling of many small catalogs. A code specially optimized for covariance computation could improve on this.

The computational cost of the LC covariance is independent of M. Thus the relative gain with respect to the sample covariance increases with increasing size of the random catalog. Since the cost of the covariance computation exceeds that of the actual 2PCF estimation by orders of magnitude, one might want to spend a bit more resources on obtaining a more accurate 2PCF estimate with a higher M, as the cost of the covariance is unaffected.

The LC covariance estimate is readily extrapolated to an arbitrary value of M, including the limit $M \to \infty$. We thus have an estimate of the error budget for any number of random points. which is valuable information when planning for an experiment.

At very small distances our method becomes less reliable due to the small number of objects in a bin. At those small distances we recommend resorting to sample covariance, which at small distances is cheap anyway.

Unlike the sample covariance, the LC covariance is not by construction positive-definite. For applications that require the covariance inverse, we recommend verifying the eigenspectrum of the constructed matrix, and for instance rebinning the data to wider bins, should the matrix turn out to be nonpositive definite.

We further derived a covariance for the estimation error of the LC covariance, and showed that it can be constructed from the components of the covariance itself. Thus, along with the covariance one can readily obtain an estimate of its errors and their correlation. We also discussed the impact on maximum-likelihood parameter estimation.

In the LC method, the covariance is estimated for small random catalogs of size $M = M_a$ and $M = 2M_a$, and the covariance for arbitrary M is constructed as a linear combination of these. We obtain the maximal reduction in the computational cost with $M_a = 1$, and we adopted this value in our validation tests. The increased uncertainty in the covariance estimate is compensated for with a larger number of catalog realizations. We have assumed that the mock catalogs are cheap to generate, so that they do not significantly contribute to the total CPU budget. Should this not be the case, the gain from the LC covariance is reduced in comparison with the sample covariance. In this case it may become beneficial to select $M_a > 1$, to reduce the variance of the covariance estimate. The selection of the optimal method is a trade-off between the cost of the 2PCF computation, the required level of accuracy, and the cost of constructing the mock catalogs.

Future large galaxy surveys, such as the one provided by *Euclid*, face the challenge of constructing the covariance for huge galaxy samples. We believe the methodology presented here provides a useful tool for meeting that challenge.

Acknowledgements. 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. The Euclid Consortium acknowledges the European Space Agency and a number of agencies and institutes that have supported the development of *Euclid*, in particular the Academy of Finland, the Agenzia Spaziale Italiana, the Belgian Science Policy, the Canadian Euclid Consortium, the French Centre National d'Etudes Spatiales, the Deutsches Zentrum für Luft- und Raumfahrt, the Danish Space Research Institute, the Fundação para a Ciência e a Tecnologia, the Ministerio de Economia y Competitividad, the National Aeronautics and Space Administration, the National Astronomical Observatory of Japan, the Netherlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, the Romanian Space Agency, the State Secretariat for Education, Research and Innovation (SERI) at the Swiss Space Office (SSO), and the United Kingdom Space Agency. A complete and detailed list is available on the *Euclid* web site (http://www.euclid-ec.org).

References

Akeson, R., Armus, L., Bachelet, E., et al. 2019, ArXiv e-prints [arXiv:1902.05569]
Alam, S., Ata, M., Bailey, S., Beutler, F., et al. 2005, MNRAS, 470, 2617
Alam, S., Aubert, M., Avila, S., et al. 2021, Phys. Rev. D, 103, 083533
Anderson, T. W. 2003, An Introduction to Multivariate Statistical Analysis (Wiley Interscience), 3rd ed.
Colavincenzo, M., Sefusatti, E., Monaco, P., et al. 2019, MNRAS, 482, 4883
Cole, S., Percival, W., Peacock, J., et al. 2005, MNRAS, 362, 505
Dávila-Kurbán, F., Sánchez, A. G., Lares, M., & Ruiz, A. N. 2021, MNRAS, 506, 4667
DESI Collaboration (Aghamousa, A., et al.) 2016, ArXiv e-prints [arXiv:1611.00036]
Dodelson, S., & Schneider, D. 2013, Phys. Rev. D, 88, 063537
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, ApJ, 633, 560
Euclid Collaboration (Scaramella, R., et al.) 2022, A&A, 662, A112
Friedrich, O., & Eifler, T. 2018, MNRAS, 473, 4150
Gaztañaga, E., & Scoccimarro, R. 2005, MNRAS, 361, 824
Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jasche, J., & Lavaux, G. 2017, A&A, 606, A37
Joachimi, B. 2017, MNRAS, 466, L83
Kalus, B., Percival, W. J., Bacon, D. J., et al. 2019, MNRAS, 482, 453
Keihänen, E., Kurki-Suonio, H., Lindholm, V., et al. 2019, A&A, 631, A73
Kitaura, F.-S., Rodríguez-Torres, S., Chuang, C.-H., et al. 2016, MNRAS, 456, 4156
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]
Manera, M., Scoccimarro, R., Percival, W. J., et al. 2013, MNRAS, 428, 1036
Merz, G., Rezaie, M., Seo, H.-J., et al. 2021, MNRAS, 506, 2503
Monaco, P. 2016, Galaxies, 4, 53
Monaco, P., Theuns, T., & Taffoni, G. 2002, MNRAS, 331, 587
Monaco, P., Di Dio, E., & Sefusatti, E. 2019, JCAP, 2019, 023
Munari, E., Monaco, P., Sefusatti, E., et al. 2017, MNRAS, 465, 4658
Padmanabhan, N., White, M., Zhou, H. H., & O'Connell, R. 2016, MNRAS, 460, 1567
Paz, D. J., & Sánchez, A. G. 2015, MNRAS, 454, 4326
Percival, W. J., Ross, A. J., Sánchez, A. G., et al. 2014, MNRAS, 439, 2531
Percival, W. J., Friedrich, O., Sellentin, E., & Heavens, A. 2022, MNRAS, 510, 3207
Pope, A. C., & Szapudi, I. 2008, MNRAS, 389, 766
Sellentin, E., & Heavens, A. F. 2016, MNRAS, 456, L132
Taylor, A., & Joachimi, B. 2014, MNRAS, 442, 2728
Taylor, A., Joachimi, B., & Kitching, T. 2013, MNRAS, 432, 1928

[1] Department of Physics, PO Box 64, 00014 University of Helsinki, Finland
e-mail: elina.keihanen@helsinki.fi

[2] Department of Physics and Helsinki Institute of Physics, Gustaf Hällströmin katu 2, 00014 University of Helsinki, Finland
[3] Dipartimento di Fisica – Sezione di Astronomia, Universitá di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
[4] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy
[5] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34143 Trieste, Italy
[6] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste TS, Italy
[7] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Str. 1, 85741 Garching, Germany
[8] INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
[9] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
[10] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[11] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK
[12] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
[13] Dipartimento di Fisica e Astronomia "Augusto Righi" – Alma Mater Studiorum Università di Bologna, via Piero Gobetti 93/2, 40129 Bologna, Italy
[14] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[15] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[16] Dipartimento di Fisica, Universitá degli studi di Genova, and INFN-Sezione di Genova, via Dodecaneso 33, 16146 Genova, Italy
[17] INFN-Sezione di Roma Tre, Via della Vasca Navale 84, 00146 Roma, Italy
[18] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[19] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
[20] Dipartimento di Fisica, Universitá degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
[21] INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
[22] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain
[23] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain
[24] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
[25] INFN section of Naples, Via Cinthia 6, 80126 Napoli, Italy
[26] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126 Napoli, Italy
[27] Dipartimento di Fisica e Astronomia "Augusto Righi" – Alma Mater Studiorum Universitá di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[28] INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy
[29] Centre National d'Etudes Spatiales, Toulouse, France
[30] Institut national de physique nucléaire et de physique des particules, 3 rue Michel-Ange, 75794 Paris Cedex 16, France
[31] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[32] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
[33] ESAC/ESA, Camino Bajo del Castillo, s/n., Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
[34] Univ Lyon, Univ Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
[35] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[36] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, 1749-016 Lisboa, Portugal

[37] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[38] Department of Astronomy, University of Geneva, ch. dÉcogia 16, 1290 Versoix, Switzerland
[39] Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, 91405 Orsay, France
[40] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
[41] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
[42] AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, 91191 Gif-sur-Yvette, France
[43] Istituto Nazionale di Astrofisica (INAF) – Osservatorio di Astrofisica e Scienza dello Spazio (OAS), Via Gobetti 93/3, 40127 Bologna, Italy
[44] Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Via Irnerio 46, 40126 Bologna, Italy
[45] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, 35122 Padova, Italy
[46] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany
[47] Dipartimento di Fisica "Aldo Pontremoli", Universitá degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
[48] INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
[49] INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
[50] Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029 Blindern, 0315 Oslo, Norway
[51] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[52] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[53] von Hoerner & Sulger GmbH, SchloßPlatz 8, 68723 Schwetzingen, Germany
[54] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
[55] Aix-Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France
[56] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, 1211 Genève 4, Switzerland
[57] NOVA optical infrared instrumentation group at ASTRON, Oude Hoogeveensedijk 4, 7991PD Dwingeloo, The Netherlands
[58] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[59] Department of Physics, Institute for Computational Cosmology, Durham University, South Road, DH1 3LE Durham, UK
[60] University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland
[61] INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
[62] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[63] European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
[64] Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, 8000 Aarhus C, Denmark
[65] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, Astrophysique, Instrumentation et Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France
[66] Space Science Data Center, Italian Space Agency, via del Politecnico snc, 00133 Roma, Italy
[67] Institute of Space Science, Bucharest 077125, Romania
[68] Dipartimento di Fisica e Astronomia "G.Galilei", Universitá di Padova, Via Marzolo 8, 35131 Padova, Italy
[69] Aix-Marseille Univ, CNRS, CNES, LAM, Marseille, France
[70] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
[71] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal
[72] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, 30202 Cartagena, Spain
[73] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
[74] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA