KiDS+VIKING-450: an internal-consistency test for cosmic shear tomography with a colour-based split of source galaxies

Li, S.-S.; Kuijken, K.; Hoekstra, H.; Hildebrandt, H.; Joachimi, B.; Kannawadi, A.

Astronomy & Astrophysics manuscript no. main
January 18, 2021

KiDS+VIKING-450: An internal-consistency test for cosmic shear tomography with a colour-based split of source galaxies

Shun-Sheng Li[1], Konrad Kuijken[1], Henk Hoekstra[1],
Hendrik Hildebrandt[2], Benjamin Joachimi[3], and Arun Kannawadi[4]

[1] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, the Netherlands
e-mail: ssli@strw.leidenuniv.nl
[2] Ruhr-Universität Bochum, Astronomisches Institut, German Centre for Cosmological Lensing (GCCL), Universitätsstr. 150, 44801 Bochum, Germany
[3] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[4] Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA

Received 25 August 2020 / Accepted 21 December 2020

ABSTRACT

We performed an internal-consistency test of the KiDS+VIKING-450 (KV450) cosmic shear analysis with a colour-based split of source galaxies. Utilising the same measurements and calibrations for both sub-samples, we inspected the characteristics of the shear measurements and the performance of the calibration pipelines. On the modelling side, we examined the observational nuisance parameters, specifically those for the redshift calibration and intrinsic alignments, using a Bayesian analysis with dedicated test parameters. We verified that the current nuisance parameters are sufficient for the KV450 data to capture residual systematics, with slight deviations seen in the second and the third redshift tomographic bins. Our test also showcases the degeneracy between the inferred amplitude of intrinsic alignments and the redshift uncertainties in low redshift tomographic bins. The test is rather insensitive to the background cosmology and, therefore, can be implemented before any cosmological inference is made.

Key words. cosmology: observations – gravitational lensing: weak – method: statistical – surveys

1. Introduction

Cosmic shear, the coherent distortion of distant galaxy shapes that arises from weak gravitational lensing by large-scale structures, is sensitive to the amplitude of matter density fluctuations, which are usually quantified by σ_8[1] and to the mean matter density Ω_m. Therefore, the main result from a cosmic shear survey is conventionally reported as a derived parameter $S_8 \equiv \sigma_8\,(\Omega_\mathrm{m}/0.3)^{0.5}$. Alternatively, the cosmic microwave background (CMB) measurements can infer the local density fluctuations by extrapolating the measured amplitude of temperature fluctuations at recombination, assuming a cosmological model. Hence, by comparing the results from these two different probes, we can test the cosmological models.

As for the current standard model of cosmology, dubbed Λ cold dark matter (ΛCDM), the latest results from *Planck* (Planck Collaboration et al. 2018) yield a constraint of $S_8 = 0.832 \pm 0.013$ (68% credible region), which is slightly higher than the results from the recent cosmic shear surveys, such as the Dark Energy Survey (DES; Troxel et al. 2018, $S_8 = 0.782^{+0.027}_{-0.027}$), the Hyper Suprime-Cam Subaru Strategic Program (HSC; Hikage et al. 2019, $S_8 = 0.780^{+0.030}_{-0.033}$), and especially the Kilo-Degree Survey (KiDS; Hildebrandt et al. 2020, hereafter H20, $S_8 = 0.737^{+0.040}_{-0.036}$).

In the era of 'precision cosmology', we have to be careful about any potential systematic effects associated with observations when interpreting results from different surveys. Given this consideration, performing internal-consistency checks is a standard part of any cosmological probe. A cosmic shear study typically bases its internal-consistency tests on a split of the estimated two-point shear correlations (Köhlinger et al. 2019; or Sect. 7.4 of H20). By assigning duplicated model parameters to each subset, one can perform theoretical modelling of the reconstructed data vector and quantify the data consistency by comparing the duplicated model parameters. This approach is useful to check for potential inconsistencies for a specific sample of source galaxies. However, the robustness is only tested at a late stage of the analysis, whilst doubling cosmological parameters comes at a considerable computational cost. The latter prevents further splits of the source sample in practice, whereas such splits can be particularly interesting because the systematics may differ.

Source galaxy properties challenge the calibration pipelines in mainly two ways: the shape measurements and the redshift estimates. First, different galaxy samples usually have different distributions of ellipticities, with red, early-type galaxies tending to have rounder shapes than their blue, late-type counterparts (Hill et al. 2019; Kannawadi et al. 2019, hereafter K19). This introduces a correlation between the shear bias and underlying galaxy sample, mainly because the shape measurements are sensitive to the distributions of galaxy ellipticities, for example, the *lens*fit algorithm used in the KiDS survey assigns weights to the measured ellipticities, resulting in a bias towards intermediate ellipticity values (Fenech Conti et al. 2017). Second, both the accuracy and the precision of a photometric redshift estimate depends on broad spectral features of a galaxy, for example, the

[1] The standard deviation of linear-theory density fluctuations in a sphere of radius $8h^{-1}$ Mpc, where $H_0 = 100h$ km s^{-1} Mpc^{-1}.

Balmer break below 4000Å (Salvato et al. 2019). The significance of these broad spectral features varies by galaxy spectral types. Generally speaking, galaxies with an old stellar population appear red at rest-frame optical wavelengths and have a pronounced 4000Å break. The bluer the galaxy, the more young stars it contains, washing out the Balmer break and other broad spectral features. Therefore, the error in photometric redshifts correlates with the galaxy spectral type (Mo et al. 2010).

We consider these sample-related systematic effects, specifically the photometric redshift uncertainty, in the KiDS cosmic shear analysis. We split the source galaxies into two mutually exclusive sub-samples according to their spectral types and apply the same measurement and calibration pipelines to these two sub-samples. This way we explored how sample-related systematics can alter the measurements and how well the calibration pipelines can assuage these effects. This split also has implications for the modelling of intrinsic alignments, which have to be taken into account explicitly. To quantify the consistency, we performed a Bayesian analysis with dedicated test parameters describing relative deviations of the nuisance parameters between the two sub-samples. By checking their posterior distributions, we can verify if the original setting suffices to capture the residual biases. The analysis code is publicly available[2].

Our approach complements other studies that check for consistency in the inferred cosmological parameters by removing tomographic bins (Köhlinger et al. 2019), or by splitting the sample by galaxy type (Samuroff et al. 2019), whilst marginalising over the corresponding nuisance parameters. We explored a different aspect: we fixed cosmological parameters but explored changes in the nuisance parameters instead. We found that our approach can test for inconsistencies in the redshift distributions and highlights the degeneracy between the redshift uncertainties and the apparent intrinsic alignment signals in a cosmology-insensitive fashion.

This paper is organised as follows. In Sect. 2, we briefly describe the cosmic shear catalogues under consideration. We show the redshift calibration in Sect. 3 and the shear bias calibration in Sect. 4. We then introduce measuring and modelling the shear signal in Sect. 5. We discuss the covariance matrix and the consistency tests in Sect. 6. The main results are presented in Sect. 7, and we summarise in Sect. 8.

2. Data

Our test is based on the first release of optical+infrared KiDS cosmic shear data dubbed KiDS+VIKING-450 (KV450; Wright et al. 2019, hereafter W19)[3]. It includes four-band optical photometry (*ugri*) from the first three data releases of KiDS (de Jong et al. 2015, 2017) and five-band near-infrared photometry ($ZYJHK_s$) from the overlapping VISTA Kilo-Degree Infrared Galaxy Survey (VIKING, Edge et al. 2013).

Details on the derivation and verification of these cosmic shear catalogues can be found in the main KiDS cosmic shear papers (Hildebrandt et al. 2017; H20) and their companion papers (Fenech Conti et al. 2017; W19). For reference, the public catalogues contain all of the necessary information to conduct a tomographic cosmic shear analysis. Amongst the most important columns are the photometric redshifts (photo-z, or z_B as in the catalogues) and the galaxy shapes (described by two ellipticity components ϵ_1, ϵ_2). The z_B values are estimated using the Bayesian photometric redshift code (BPZ; Benítez 2000; Coe et al. 2006) with an improved redshift prior from Raichoor et al. (2014) and the nine-band photometry from W19. The galaxy shapes are measured from the r-band images (median seeing 0.7″) using the *lens*fit algorithm (Miller et al. 2007; Kitching et al. 2008; Miller et al. 2013) with a self-calibration for noise bias (Fenech Conti et al. 2017).

[2] https://github.com/lshuns/CosmicShearRB
[3] http://kids.strw.leidenuniv.nl/DR3/kv450data.php

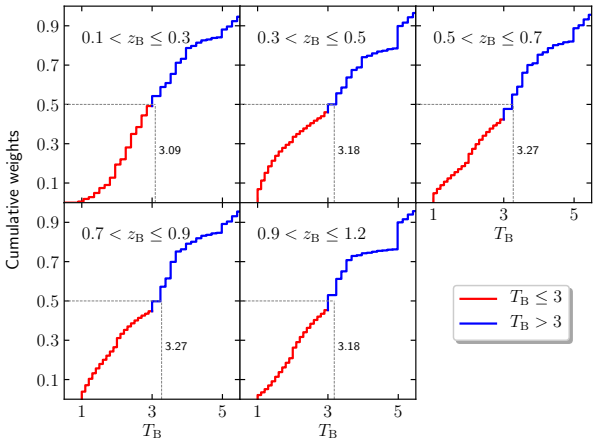

Fig. 1. Cumulative *lens*fit-weighted distributions of T_B values. The dashed line indicates the ideal half-half split in each tomographic bin, which is close to our split at $T_B = 3$.

Throughout this study, we only used sources with valid nine-band photometry (`GAAP_Flag_ugriZYJHKs==0`). This mask reduces the original area by ∼ 5% and retains ∼ 13 million objects, which is identical to the choice made by the main KV450 cosmic shear analysis. Following H20, we binned source galaxies into five tomographic bins defined as $0.1 < z_B \leq 0.3$, $0.3 < z_B \leq 0.5$, $0.5 < z_B \leq 0.7$, $0.7 < z_B \leq 0.9$, $0.9 < z_B \leq 1.2$. Given the purposes of checking systematic effects caused by galaxy properties, we further split the whole sample into two sub-samples according to the spectral types of source galaxies. This is achieved by using the T_B values reported by the BPZ code during the photo-z estimating procedure (see Benítez 2000, for a detailed discussion). Briefly, the T_B value is calculated within a Bayesian framework using six templates of galaxy spectra (Coleman et al. 1980; Kinney et al. 1996). We defined our two sub-samples as $T_B \leq 3$ (a combination of E1, Sbc, Scd types, labelled as 'red' in this paper) and $T_B > 3$ (a combination of Im and two starbust types, labelled as 'blue' in this paper). This cut is chosen to ensure similar statistical power in the two sub-samples (see Fig. 1). Source properties of these two sub-samples are summarised in Table 1.

3. Calibration of redshift distributions

One of the most challenging tasks for a tomographic cosmic shear study is to estimate the source redshift distribution for each tomographic bin. These intrinsic redshift distributions vary with galaxy samples, so we need to calibrate the photo-z estimates in the two sub-samples, separately. We followed the fiducial technique, dubbed DIR in H20, for this task. This method directly estimates the underlying redshift distributions of a photometric sample using deep spectroscopic redshift (spec-z) catalogues that overlap with the photometric survey. We shortly discuss our implementation of this method in this section and refer interested

Table 1. Source information in the two sub-samples.

Sample	Bin	Photo-z range	Total *lens*fit weights	$n_{\rm eff}$ $\left(\mathrm{arcmin}^{-2}\right)$	$\sigma_{\epsilon,i}$	m-bias	Mean($z_{\rm DIR}$)	Median($z_{\rm DIR}$)
$T_{\rm B} \leq 3$	1	$0.1 < z_{\rm B} \leq 0.3$	7,031,963	0.38	0.279	-0.029 ± 0.010	0.351	0.282
(red)	2	$0.3 < z_{\rm B} \leq 0.5$	10,404,223	0.59	0.252	-0.009 ± 0.007	0.430	0.396
	3	$0.5 < z_{\rm B} \leq 0.7$	15,508,696	0.90	0.276	-0.010 ± 0.007	0.546	0.531
	4	$0.7 < z_{\rm B} \leq 0.9$	9,837,460	0.64	0.250	0.008 ± 0.006	0.744	0.732
	5	$0.9 < z_{\rm B} \leq 1.2$	8,466,542	0.59	0.275	0.006 ± 0.008	0.909	0.894
$T_{\rm B} > 3$	1	$0.1 < z_{\rm B} \leq 0.3$	7,269,125	0.42	0.270	-0.004 ± 0.008	0.437	0.244
(blue)	2	$0.3 < z_{\rm B} \leq 0.5$	12,200,673	0.75	0.277	-0.007 ± 0.006	0.573	0.431
	3	$0.5 < z_{\rm B} \leq 0.7$	21,116,034	1.46	0.292	-0.002 ± 0.006	0.791	0.644
	4	$0.7 < z_{\rm B} \leq 0.9$	12,134,896	0.92	0.286	0.026 ± 0.006	0.914	0.842
	5	$0.9 < z_{\rm B} \leq 1.2$	10,207,426	0.87	0.293	0.036 ± 0.009	1.081	1.022

Notes. The effective number density $n_{\rm eff}$ is calculated from Eq. (1) of Heymans et al. (2012a). The reported ellipticity dispersion is defined as $\sigma_{\epsilon,i} = (\sigma_{\epsilon 1} + \sigma_{\epsilon 2})/2$. The m-bias is defined in Eq. (1) and detailed in Sect. 4. Reported uncertainties were computed from the dispersion of 50 bootstrap samples. The mean and median of the redshift distributions were obtained from the DIR calibration, which is detailed in Sect. 3.

readers back to the original papers for more details (Lima et al. 2008; Hildebrandt et al. 2017, 2020).

The DIR method requires that the calibration sample (the spec-z sample) spans, at least sparsely, the full extent of the multi-band magnitude space covered by the target sample (the photo-z sample) and that the mapping from magnitude space to redshift space is unique. Therefore, the coverage of the spec-z sample is essential for the accuracy of this method. We here used the same set of spec-z catalogues as used in the fiducial KV450 cosmic shear analysis. It includes the zCOSMOS survey (Lilly et al. 2009), the DEEP2 survey (Newman et al. 2013), the VIMOS VLT Deep survey (Le Fèvre et al. 2013), the GAMA-G15Deep survey (Kafle et al. 2018) and a combined catalogue provided by ESO in the *Chandra* Deep Field South area[4]. These independent spec-z surveys with different lines-of-sight and depths minimise shot noise and sample variance in the calibration sample.

Since the spec-z catalogues cannot fully represent the photometric sample, one needs to weight spec-z objects to ensure a suitable match between the spectroscopic and photometric distributions. The method, based on a kth nearest neighbour (kNN) approach, is detailed in Sect. 3 of Hildebrandt et al. (2017). Briefly, it assigns weights to the spec-z objects by comparing the volume densities of the spec-z and photometric objects in the nine-band magnitude space ($ugriZYJHK_s$). Therefore, KiDS+VIKING-like observations are required in the same areas as the aforementioned spec-z surveys. H20 have built these photometric observations from multiple ways given the availability of specific data sets in those spec-z survey fields. We adopted the same sample and split it with the same criterion as used for the main KV450 sample to build two representatives of our two sub-samples.

The resulting redshift distributions of the two sub-samples are shown in Fig. 2. Also presented are the mean and median differences between these two redshift distributions (see Table 1 for separate values). The importance of photo-z calibration is demonstrated by the tails of the DIR redshift distributions compared to the ranges selected by the photo-z cuts (shaded regions). These differences between the DIR results and photo-z estimates are more significant in the red sub-sample, where an overall bias towards overestimating photo-z is shown. This may seem counterintuitive at first given the discussion presented in Sect. 1, which states that young stars can wash out spectral features for photo-z estimation resulting in larger errors in bluer galaxies. However, we stress that the red sub-sample defined in Sect. 2 is not 'purely red', but also includes Sbc and Scd types (see Sect. 2), which could worsen the photo-z estimates. For our purposes, we are interested in the redshift difference between the two sub-samples. As can be seen, the differences are significant with the median differences as high as ~ 0.13 and the mean differences ~ 0.24 in certain bins. This level of difference will result in considerably different cosmic shear signals for the two sub-samples (see Sect. 5).

In practice, the DIR method is susceptible to various systematic effects, mainly induced by the incompleteness of the spec-z sample, due to selection effects and sample variance in the different spectroscopic surveys that make up the spec-z catalogue (see Wright et al. 2020a, for an updated method that is more robust to such incompleteness). To account for these potential systematic effects, H20 introduced five nuisance parameters δ_{z_i} in their model to allow for linear shifts of the redshift distributions $n_i(z) \rightarrow n_i(z + \delta_{z_i})$ (see Table. 2). Priors for these parameters are obtained using a spatial bootstrapping approach. In our consistency tests described below we focus on an extension of these nuisance parameters to the colour-split sub-samples (see Sect. 6).

4. Calibration of shape measurements

The shape measurements are susceptible to various biases due to the noise of galaxy images, the complexity of galaxy shapes, the selection effects and so on (see Sect. 2 of K19, for a theoretical discussion). The weak lensing community have performed several blind challenges to test the performance of shape measurement pipelines (Heymans et al. 2006; Massey et al. 2007; Bridle et al. 2010; Kitching et al. 2012; Mandelbaum et al. 2015). These tests, based on simplified image simulations, are useful to understand common sources of shear bias, but cannot eliminate biases in a specific survey. In particular, differences in selection criteria between surveys affect the shear bias. These residual biases need to be calibrated with dedicated, tailor-made image simulations (Hoekstra et al. 2015). Following Heymans et al. (2006), we quantify these residual biases using a linear parameterisation

$$g_i^{\rm obs} = (1 + m_i) g_i^{\rm true} + c_i \,, \qquad (1)$$

[4] http://www.eso.org/sci/activities/garching/projects/goods/MasterSpectroscopy.html

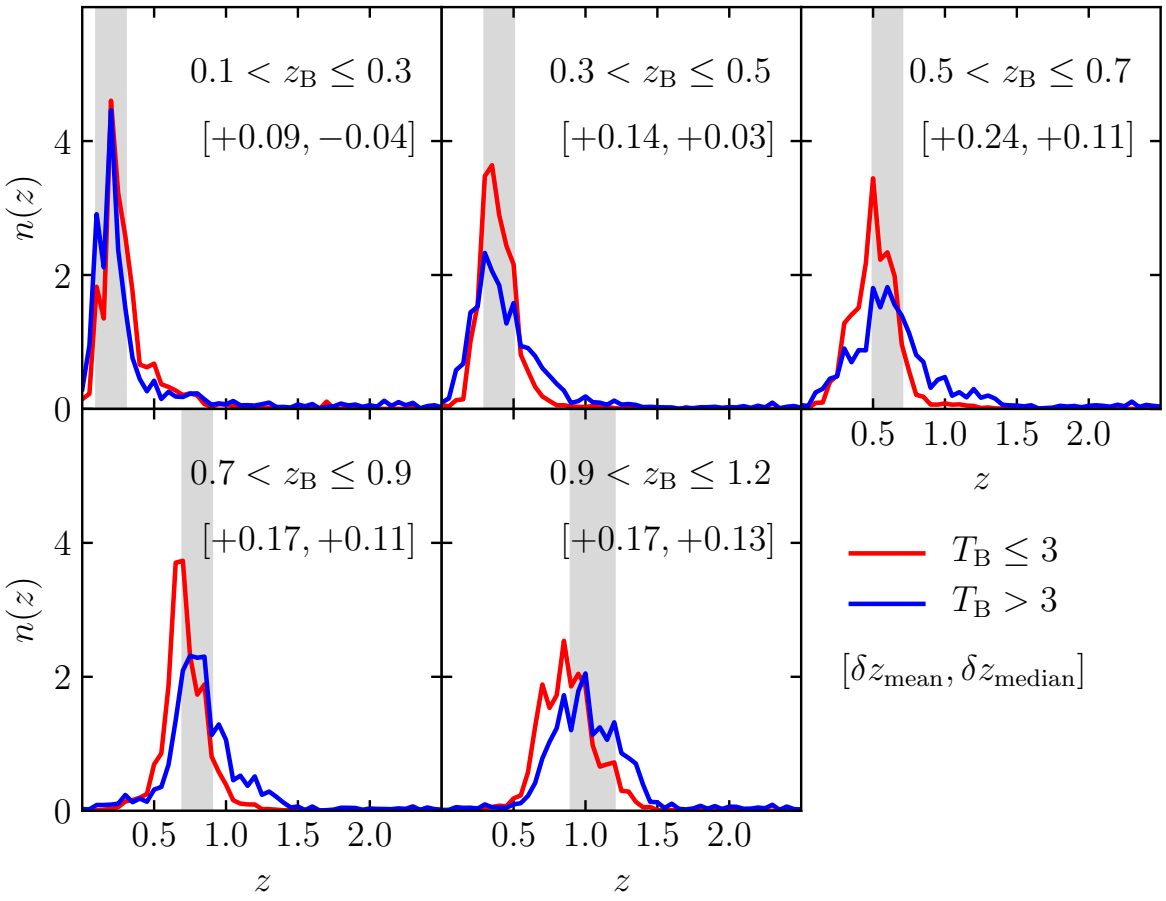

Fig. 2. Redshift distributions for the two sub-samples, estimated from DIR technique. Shaded regions correspond to photo-z cuts for the tomographic binning. Mean and median differences were calculated as $\delta z_{\rm mean/median} = z_{\rm mean/median,blue} - z_{\rm mean/median,red}$.

where $g_i^{\rm obs}$ and $g_i^{\rm true}$ are the observed and the true gravitational shears, respectively, with $i = 1, 2$ referring to the two different components. In practice, we found isotropy of m results, that is $m_1 \approx m_2$, so we simply adopt $m = (m_1 + m_2)/2$.

The two types of biases m (the multiplicative bias) and c (the additive bias or c-term) have different sources and properties. The former is usually determined from image simulations, whereas the latter can be inferred directly from the data. As K19 show, shear biases depend not only on the selection function but also on the overall population of the galaxies. Therefore shear calibrations should be performed separately for samples containing different galaxy populations. This was the case for the different tomographic bins in the KV450 analysis and applies even more so to our split analysis.

We therefore re-estimated multiplicative biases in the two sub-samples using the COllege simulations (COSMOS-like lensing emulation of ground experiments, K19), which were also used in the current KV450 cosmic shear analysis. The main features of the COllege simulations are the observation-based input catalogue and the assignment of photometric redshifts. The input catalogue contains information on galaxy morphology and position from *Hubble* Space Telescope observations (Griffith et al. 2012) of the COSMOS field (Scoville et al. 2007). The photometric redshifts of simulated galaxies are assigned by cross-matching the input catalogue to the KiDS catalogue. This setup ensures a high level of realism of the simulated catalogue and allows us to analyse the simulated data using the same pipelines as for the real data. K19 have demonstrated that the simulated catalogue matches the full KV450 catalogue faithfully in all crucial properties including the galaxy shapes, sizes and positions.

As expected, we found noticeable differences in the galaxy properties for the two sub-samples. We demonstrate one of these comparisons in Fig. 3, which compares the distributions of galaxy ellipticities. As already mentioned in Sect. 1, the ellipticity variance is one of the main sources of shape measurement biases (see also Viola et al. 2014) and therefore an indication of the variance of shear biases in the two sub-samples.

Our calibration approach is identical to that used in the fiducial KV450 cosmic shear analysis. It adopts a re-weighting scheme named as "Method C" in Fenech Conti et al. (2017) to account for slight differences between the observations and the simulations. The m value is reported per tomographic bin using a weighted average of individual galaxies belonging to the corresponding tomographic bin. We refer readers to Sect. 6 of K19 for details.

We show our estimates of multiplicative biases for the two sub-samples in Fig. 4, compared with the results from the whole sample. The five sections from top to bottom correspond to the

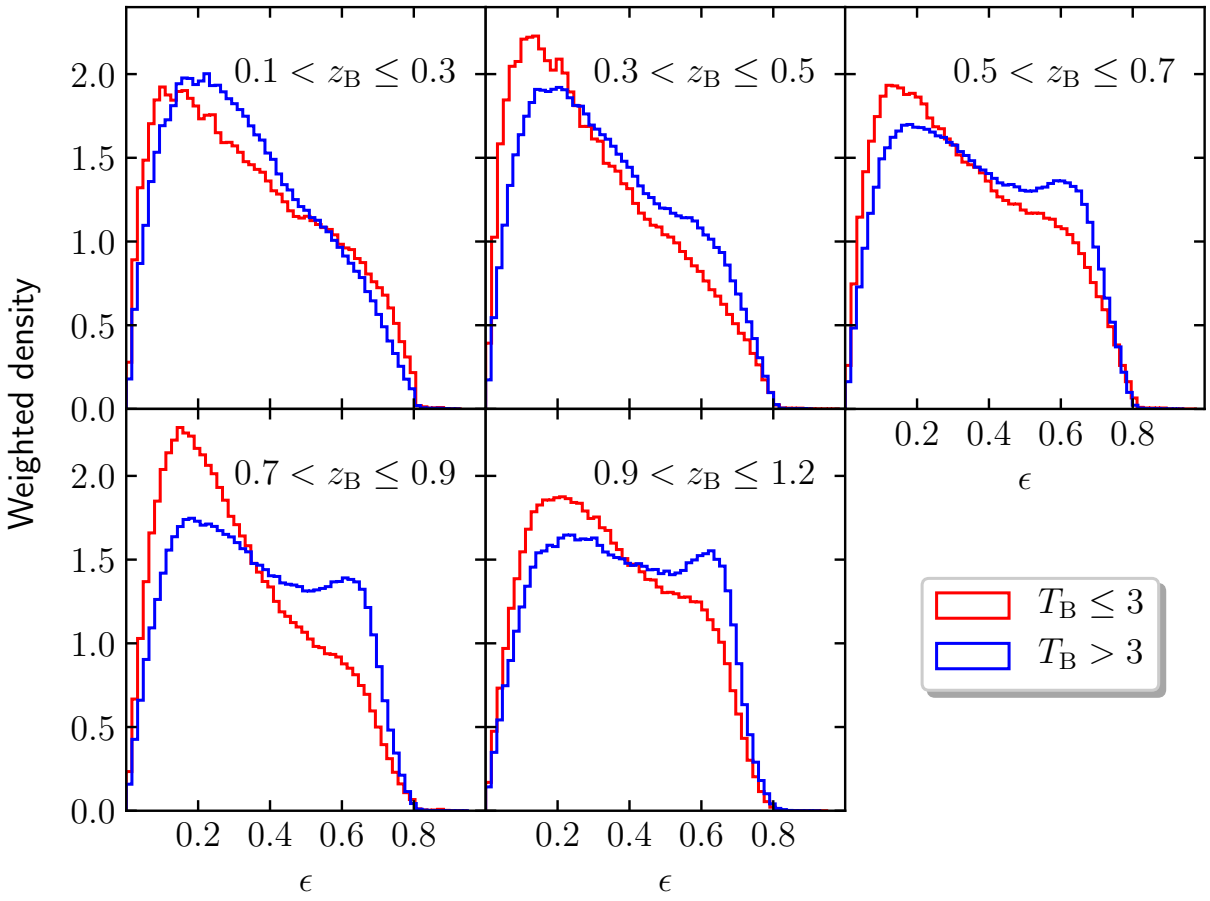

Fig. 3. Normalised *lens*fit-weighted distributions of ellipticities of galaxies in the two sub-samples. The ellipticity is defined as $\epsilon = \sqrt{\epsilon_1^2 + \epsilon_2^2}$. We note that the different distributions reflect different galaxy populations and indicate different shear biases in the two sub-samples.

five tomographic bins from lower to higher redshifts. We noticed some significant differences in the m values, especially for higher tomographic bins: these are mainly caused by the differences in the ellipticity distributions presented in Fig. 3. However, when considering the impact on the cosmic shear signals, the adjustments induced by these m-value differences are much smaller than those caused by the redshift differences as demonstrated in Fig. 5. We thus assumed that residual systematics from the shear calibration are secondary and focus our consistency tests on the redshift calibration.

The treatment of additive bias is sophisticated in the fiducial KV450 cosmic shear analysis (see Sect. 4 of H20, for details). Briefly, the treatment can be summarised as three aspects: First, the value of c_i in each tomographic bin and in each patch is estimated by averaging over the measured galaxy ellipticities. These c_i values are then subtracted from the galaxy ellipticities before the shear correlation functions are calculated (Eq. 2). Second, a nuisance parameter δ_c is introduced into the model to account for a potential offset of the empirically determined c_i values. The result from forward-modelling suggests that δ_c is very close to 0 (see Table 2). Third, a position-dependent additive bias pattern in the ϵ_1 ellipticity component is introduced to account for an imperfection in the OmegaCAM detector chain. This pattern is publicly available as a supplementary file along with the main cosmic shear catalogues. Furthermore, another nuisance parameter A_c is introduced to allow an overall scaling of this 2D pattern (see Table 2).

We mainly followed this strategy for the additive bias calibration. We corrected the c-term per tomographic bin and per patch using the same empirical approach mentioned above. We also included the 2D c-term pattern in our models. But we abandoned the two nuisance parameters δ_c and A_c from our model, as they do not have a significant impact on the fit.

5. Shear signal

The cosmic shear signal is encoded in the measured shapes of source galaxies as small coherent distortions. Therefore, proper statistical measures and models are required for a cosmic shear study. We detail these processes in this section. We first built the joint data vector for the two sub-samples with estimates of the shear correlation functions in Sect. 5.1 and then modelled it taking various astrophysical and cosmological effects into account in Sect. 5.2. The setup is based on the fiducial analysis of H20 but with slight adjustments to meet our test purpose.

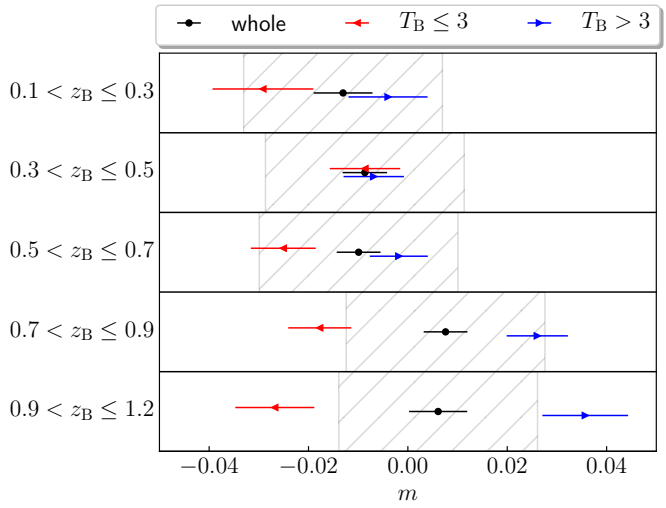

Fig. 4. Multiplicative biases for the two sub-samples and the whole sample in each tomographic bin. Errors shown were estimated from bootstrapping. The hatched regions indicate the 0.02 error budget adopted by H20.

5.1. Statistical measures

The shear signal is captured by two-point shear correlation functions, which can be estimated from two tomographic bins i and j as

$$\xi_{\pm}^{ij}(\theta) = \frac{\sum_{ab} w_a w_b \left[\epsilon_t^i(\boldsymbol{x}_a)\epsilon_t^j(\boldsymbol{y}_b) \pm \epsilon_\times^i(\boldsymbol{x}_a)\epsilon_\times^j(\boldsymbol{y}_b)\right]}{(1+m^i)(1+m^j)\sum_{ab} w_a w_b} , \tag{2}$$

where $\epsilon_{t,\times}$ are the tangential and cross ellipticities regarding the vector $\boldsymbol{x}_a - \boldsymbol{y}_b$ between a pair of galaxies (a, b), and w is the *lens*fit weight. The summation runs over all galaxy pairs within an assigned spatial bin $\Delta\theta$ for each $\theta = |\boldsymbol{\theta}_b - \boldsymbol{\theta}_a|$. The multiplicative biases m^i were obtained in Sect. 4 for each tomographic bin i.

We calculated Eq. (2) for the two sub-samples, separately, using the public TREECORR code[5] (Jarvis et al. 2004). The spatial binning is identical to that used in H20, that is, nine logarithmically spaced bins within the interval $[0.5', 300']$. We used the first seven bins for ξ_+, and the last six bins for ξ_-. These criteria are chosen to mitigate baryon feedback on small scales and the additive shear biases on large scales (see H20, for details). The joint data vector $(\xi_\pm^{\rm blue}, \xi_\pm^{\rm red})$ we built through these measurements contains $(7 + 6) \times 15 \times 2 = 390$ points.

We show our estimates of the data vector in Fig. 5 with differences defined as $\Delta\xi_\pm = \xi_\pm^{\rm blue} - \xi_\pm^{\rm red}$. The errors shown were adopted from the analytical covariance matrix detailed in Sect. 6.1. Two series of data vectors correspond to the results with and without the multiplicative shear calibration. The difference is minor as expected given the overall small m values (see Table 1). Some non-zero trends are present in several bins, which are in principle caused by the different redshift distributions of these two sub-samples, as shown in Fig. 2. We detail how the redshift distributions can explain these measurements in the following section.

5.2. Theoretical modelling

The measured correlation functions $\xi_\pm^{ij}(\theta)$ are related to the lensing convergence power spectrum $P_\kappa^{ij}(\ell)$ through (see e.g. Bartelmann & Schneider 2001)

$$\xi_\pm^{ij}(\theta) = \frac{1}{2\pi}\int {\rm d}\ell\ \ell P_\kappa^{ij}(\ell) J_{0/4}(\ell\theta) , \tag{3}$$

where ℓ is the angular wavenumber in the Fourier domain, and $J_{0/4}(\ell\theta)$ are Bessel functions of the first kind, with J_0 denoting the zeroth-order (for ξ_+) and J_4 the fourth-order (for ξ_-). Using the Kaiser-Limber approximation (Limber 1953; Kaiser 1992, 1998; Loverde & Afshordi 2008), $P_\kappa^{ij}(\ell)$ is in turn related to the physical matter power spectrum P_δ, via

$$P_\kappa^{ij}(\ell) = \int_0^{\chi_{\rm H}} {\rm d}\chi\ \frac{q_i(\chi)q_j(\chi)}{[f_{\rm K}(\chi)]^2} P_\delta\left(\frac{\ell+1/2}{f_{\rm K}(\chi)}, \chi\right) , \tag{4}$$

where χ and $f_{\rm K}(\chi)$ are the comoving radial distance and the comoving angular distance, respectively. The upper limit of the integral $\chi_{\rm H}$ is the comoving horizon distance. The lensing efficiency $q_i(\chi)$ for tomographic bin i is defined as

$$q_i(\chi) = \frac{3H_0^2\Omega_{\rm m}}{2c^2}\frac{f_{\rm K}(\chi)}{a(\chi)}\int_\chi^{\chi_{\rm H}} {\rm d}\chi'\ n_i(\chi')\frac{f_{\rm K}(\chi'-\chi)}{f_{\rm K}(\chi')} , \tag{5}$$

which depends on the redshift distribution of galaxies $n_i(\chi){\rm d}\chi = n_i(z){\rm d}z$ along with other cosmological parameters. Therefore, different redshift distributions will cause a difference in shear signal between the two sub-samples.

We calculated the matter power spectrum using the Boltzmann-code CLASS (Blas et al. 2011) with non-linear corrections from HMCODE (Mead et al. 2016). Following H20, we assumed a ΛCDM model with five primary cosmological parameters and one parameter for baryonic feedback processes on small scales. They are the densities of cold dark matter and baryons ($\Omega_{\rm CDM}$ and $\Omega_{\rm b}$), the amplitude and the index of the scalar power spectrum ($\ln(10^{10}A_{\rm s})$, $n_{\rm s}$), the scaled Hubble parameter (h), and the amplitude of the halo mass-concentration relation (B).

For the purposes of consistency tests, it is unnecessary to explore this whole cosmological parameter space, which is the same for the two sub-samples. Therefore, we fixed aforementioned cosmological parameters to two different sets of best-fit values from KV450 (Hildebrandt et al. 2020) and *Planck* (Planck Collaboration et al. 2018) (see Table 2). In this way, we can simplify our theoretical models while checking for potential cosmological dependence.

The last piece of information needed for modelling the observed correlation functions is the intrinsic alignment (IA) of galaxies (Troxel & Ishak 2015; Joachimi et al. 2015). A common approach to make allowances for this effect is to add a "nonlinear linear" IA model into the measured shear signal (Hirata & Seljak 2004; Bridle & King 2007):

$$\hat{\xi}_\pm = \xi_\pm + \xi_\pm^{\rm II} + \xi_\pm^{\rm GI} , \tag{6}$$

where $\hat{\xi}_\pm$ and $\xi_\pm$ correspond to the measured shear signal and the pure cosmic shear signal, respectively. The IA signals are added as $\xi_\pm^{\rm II}$ ('intrinsic-intrinsic' term between the intrinsic ellipticities of nearby galaxies) and $\xi_\pm^{\rm GI}$ ('gravitational-intrinsic' term between the intrinsic ellipticity of a foreground galaxy and the shear experienced by a background galaxy). These two IA terms can be calculated using the same formula shown in Eq. (3) with power spectra

$$P_{\rm II}^{ij}(\ell) = \int_0^{\chi_{\rm H}} d\chi\ F^2(z)\frac{n_i(\chi)n_j(\chi)}{[f_{\rm K}(\chi)]^2} P_\delta\left(\frac{\ell+1/2}{f_{\rm K}(\chi)}, \chi\right) , \tag{7}$$

[5] https://github.com/rmjarvis/TreeCorr

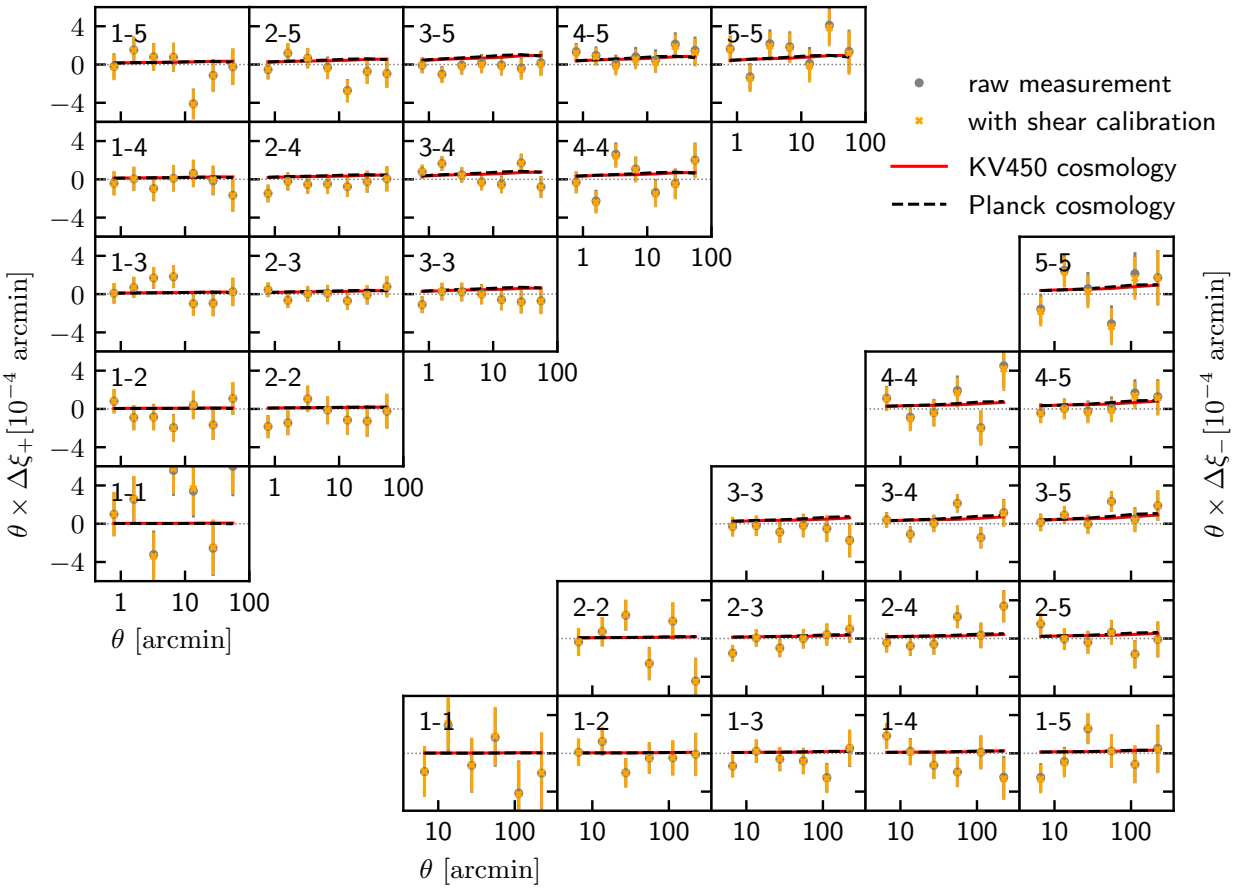

Fig. 5. Difference between two-point shear correlation functions from the two sub-samples ($\Delta\xi_\pm = \xi_\pm^{\rm blue} - \xi_\pm^{\rm red}$). The errors shown are defined as $\sigma_C = \sqrt{C_{\rm b,D} + C_{\rm r,D} - 2C_{\rm br,D}}$, where the subscript 'D' means the diagonal of a matrix, and the three unique parts of the whole covariance matrix are denoted as $C_{\rm b}$ for the blue sub-sample, $C_{\rm r}$ for the red sub-sample and $C_{\rm br}$ for their cross-covariance. We found these errors are close to the measurement errors reported by the TREECORR code ($\sigma_{\rm measure}/\sigma_C \gtrsim 0.8$), indicating that the diagonal elements of the covariance matrix are dominated by measurement noise. The overall agreement between the two sets of data vectors with and without the shear calibration (orange crosses vs. black dots) indicates the multiplicative bias has little effect in this study.

$$P_{\rm GI}^{ij}(\ell) = \int_0^{\chi_{\rm H}} d\chi\, F(z) \frac{q_i(\chi) n_j(\chi) + q_j(\chi) n_i(\chi)}{[f_{\rm K}(\chi)]^2} P_\delta\left(\frac{\ell + 1/2}{f_{\rm K}(\chi)}, \chi\right), \tag{8}$$

where

$$F(z) = -A_{\rm IA} C \rho_{\rm crit,0} \frac{\Omega_{\rm m}}{D_+(z)}. \tag{9}$$

The normalisation constant is $C = 5 \times 10^{-14} h^{-1} M_\odot^{-1} {\rm Mpc}^3$, $\rho_{\rm crit,0}$ is the critical density today, and the linear growth factor $D_+(z)$ is normalised to unity today. Following H20, we ignored the redshift and luminosity dependence of IA and leave one nuisance parameter $A_{\rm IA}$ for IA effects (but see Fortuna et al. 2020).

Now with all the information prepared, we can forward-model the shear correlation functions. For demonstration, we fixed all the model parameters and use the redshift distributions estimated in Sect. 3 to predict the joint data vector of the two sub-samples. The results are shown in Fig. 5. Two different predictions come from two different sets of cosmological parameters: the red solid line from KV450 best-fit values and the black dashed line from *Planck* best-fit values. All the other nuisance parameters are set to the best-fit KV450 results as shown in Table 2. Even with this simple setting, the predicted results generally follow the trends seen from the data, demonstrating that the redshift difference is indeed the main cause for the different shear correlation functions in the two sub-samples. The other feature worth noting is the similarity between the two predictions from the two different sets of cosmological parameters. This implies that our test model is insensitive to the background cosmology. To quantify the goodness of fit and test the robustness of the pipelines, we need a more careful Bayesian analysis with proper test models and take correlations between measurements into account.

Table 2. Model parameters and their best-fit values from KV450 cosmic shear analysis (Hildebrandt et al. 2020) and *Planck* CMB analysis (Planck Collaboration et al. 2018).

Parameter	KV450	*Planck*	Definition
$\Omega_{\rm CDM}h^2$	0.058	0.120	CDM density today
$\Omega_{\rm b}h^2$	0.022	0.022	Baryon density today
$\ln(10^{10}A_{\rm s})$	4.697	3.045	Scalar spectrum amplitude
$n_{\rm s}$	1.128	0.966	Scalar spectrum index
h	0.780	0.673	Hubble parameter
B	2.189	-	Baryon feedback amplitude
$A_{\rm IA}$	0.494	-	IA amplitude
$\delta_c \times 10^5$	2.576	-	c-term offset
A_c	1.143	-	2D c-term amplitude
δ_{z_1}	−0.006	-	Bin 1 offset
δ_{z_2}	0.001	-	Bin 2 offset
δ_{z_3}	0.026	-	Bin 3 offset
δ_{z_4}	−0.002	-	Bin 4 offset
δ_{z_5}	0.003	-	Bin 5 offset

Notes. The first five parameters are the standard cosmological parameters. Other parameters are nuisance parameters introduced by Hildebrandt et al. (2020) to account for various effects associated with cosmic shear analysis. The KV450 best-fit values are extracted from the primary Monte Carlo Markov Chain, which is publicly available at `http://kids.strw.leidenuniv.nl/cosmicshear2018.php`. The *Planck* best-fit values correspond to the TT,TE,EE+lowE+lensing results with the `Plik` likelihood (Table 1 of Planck Collaboration et al. 2018).

6. Consistency tests

Quantifying the internal consistency is not a trivial task given the correlations between measurements and the difficulty in comparing different models. On the one hand, neglecting intrinsic correlations between measurements can lead to untrustworthy conclusions. As demonstrated by Köhlinger et al. (2019), a lack of consideration of correlations can confuse residual systematics with the overall goodness of fit. On the other hand, null tests based on global summary statistics, such as Bayesian evidence, are practically difficult for high-dimensional models (see e.g. Trotta 2008). Moreover, different prior choices between hypotheses can complicate the interpretation of the final results (Handley & Lemos 2019b; Lemos et al. 2019).

We address these issues in this section. We first built an analytical covariance matrix to account for all the correlations between measurements (Sect. 6.1). We then performed a Bayesian analysis with dedicated test parameters to quantify the potential discrepancy between measurements from the two sub-samples (Sect. 6.2). The conclusion is based on the posterior distributions of these test parameters. Through this approach, we can balance accuracy and simplicity in our model.

The modelling pipeline detailed below is publicly available[6]. It is a modified version of the MontePython package (Audren et al. 2013; Brinckmann & Lesgourgues 2018) with the PyMultiNest algorithm (Buchner et al. 2014), which is a python wrapper of the nested sampling algorithm MultiNest (Feroz et al. 2009). The original MontePython package is adopted for the KV450 cosmological analysis in H20 and the consistency tests with a split of data vector (Köhlinger et al. 2019).

[6] `https://github.com/lshuns/montepython_KV450`

6.1. Covariance matrix

We estimated the covariance matrix for the joint data vector built in Sect. 5.1 using the analytical model developed in Hildebrandt et al. (2017), H20 and Joachimi et al. (2020). The analytical approach is an improvement over the usual numerical or Jackknife approach with advantages in dealing with effects from modelling the noise and the finite survey areas. We here only briefly summarise the main features of this analytical recipe and refer interested readers to Sect. 5 of Hildebrandt et al. (2017) and Joachimi et al. (2020) for details.

The analytical model comprises three terms: a Gaussian term associated with sample variance and shape noise, a non-Gaussian term from in-survey modes, and a third term, which is also non-Gaussian, from super-survey modes (known as super-sample covariance; SSC). The first, Gaussian term is estimated following Joachimi et al. (2008), with a transfer function from Eisenstein & Hu (1998) and the non-linear corrections from Takahashi et al. (2012). The source information used is listed in Table 1; these are the effective galaxy number density ($n_{\rm eff}$) and the weighted ellipticity dispersion ($\sigma_{\epsilon,i}$). The second, non-Gaussian term is calculated using the formalism from Takada & Hu (2013) with the halo mass function and halo bias from Tinker et al. (2010). The halo profile is described using a Fourier-transform version (Scoccimarro et al. 2001) of the NFW model (Navarro et al. 1996), with the concentration-mass relation from Duffy et al. (2008). The final, SSC term is again modelled using the formalism from Takada & Hu (2013), and the survey footprint is modelled with a HEALPix map (Górski et al. 2005).

The shear calibration presented in Sect. 4 also suffers from uncertainties. We adopted a systematic uncertainty $\sigma_m = 0.02$ for the multiplicative biases as estimated by K19 and used in H20 and Wright et al. (2020b) and propagated it into the covariance matrix through $C_{ij}^{\rm cal} = 4\xi_i^{\rm T}\xi_j^{\rm T}\sigma_m^2 + C_{ij}$, where $\xi^{\rm T}$ is the joint data vector predicted using the KV450 best-fit values and the DIR redshift distributions (see Sect. 3). We ignored the error of the additive biases due to its negligible effect (see Appendix D4 of Hildebrandt et al. 2017, for a detailed discussion).

We show the final correlation matrix for the joint data vector in Fig. 6. Non-negligible contributions from off-diagonal regions are easily noticed, indicating the non-trivial correlations between the measurements both within individual sub-samples and across the two sub-samples. The importance of the potential correlations between (two) parts of a split was already highlighted in Köhlinger et al. (2019), but here we confirmed it more directly. By including the full covariance matrix into our consistency tests, we naturally took all the data correlations into account.

We inspected the relative contributions of the Gaussian and non-Gaussian terms to the full covariance matrix. We found that the Gaussian term generally dominates over the non-Gaussian term in the diagonal parts, but the latter contributes more in the off-diagonal regions. This general behaviour is more clearly demonstrated in Joachimi et al. (2020). Since our test model is most sensitive to the difference $\Delta\xi$ between the two sub-samples, we constructed the covariance matrix of $\Delta\xi$ as $C_\Delta = C_{\rm blue} + C_{\rm red} - 2C_{\rm cross}$, and compared it to the covariance matrices of the single data vectors ($\xi_{\rm blue}$ or $\xi_{\rm red}$). We found that the non-Gaussian contributions are significantly suppressed in C_Δ with an overall reduction of $\lesssim 75\%$ compared to $C_{\rm blue}$. The Gaussian contributions are also slightly suppressed, mainly in the off-diagonal regions. The cancellation of sample variance can explain both suppressions in the covariance matrix C_Δ. Therefore,

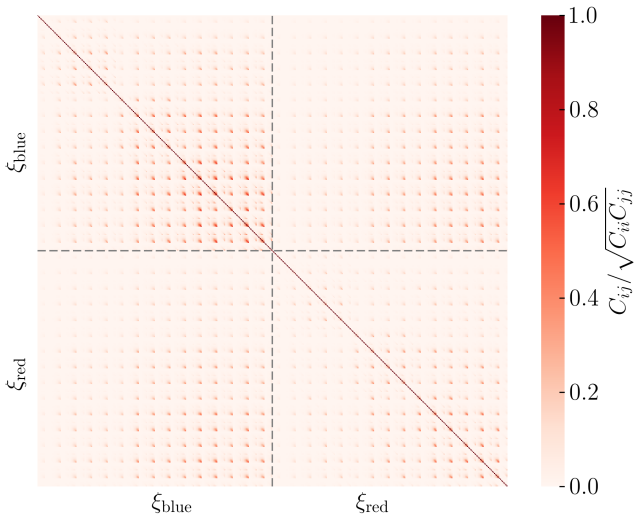

Fig. 6. Analytical correlation matrix for the joint data vector. The covariance C_{ij} is normalised using the diagonal $\sqrt{C_{ii}C_{jj}}$ to show the correlation matrix.

we verified that our test model is robust against uncertainties in the sample variance and changes in the cosmological parameters.

6.2. Test setup

With the covariance matrix prepared, we can now explore the parameter space with a Bayesian analysis. Our primary objective is to check if a common set of nuisance parameters is sufficient to capture residual systematics in the two sub-samples. For this purpose, we fixed all the cosmological parameters, which in principle share the same values in the two sub-samples. We verified this cosmology insensitivity assumption by running an additional chain with free cosmological parameters. The results are consistent with the fixed-cosmology settings, and we hardly observe any degeneracy between cosmological parameters and test parameters. In what follows, we therefore fixed the cosmology parameters to simplify the likelihood function and avoid unnecessary exploration of the high-dimensional parameter space. To account for any potential residual effects from an 'incorrect' choice of cosmological parameters, we ran two setups with cosmological parameters from the KV450 cosmic shear results and from the *Planck* CMB results (see Table 2).

We built our test model $\mathcal{H}_1$ by introducing six test parameters besides the nuisance parameters used in H20: a shift in IA amplitude $A_{\rm IA,s}$ and shifts in redshift offsets $\delta_{z_i,{\rm s}}$. We implemented them in the two sub-samples as

$$X_{\rm blue/red} = X \pm X_{\rm s} \ , \tag{10}$$

where X represents the $A_{\rm IA}$ or δ_{z_i} parameters, whereas $X_{\rm s}$ denotes corresponding test parameters. The plus sign is applied to the blue sub-sample, and the minus sign is for the red sub-samples. While a difference in the IA signal is expected, the differences in redshift offsets should vanish if the calibration pipeline is robust against sample-related systematics. Any non-vanishing values of $\delta_{z_i,{\rm s}}$ imply residual systematics that cannot be adequately captured by the common nuisance parameters. We therefore based our result mainly on the posterior distributions of these test parameters. In addition, we set up a base model $\mathcal{H}_0$ for a control purpose, where we adopted the same set of nuisance parameters as in H20 to model the joint data vector built from our two sub-samples. It includes six free nuisance parameters: the amplitude of the IA signal $A_{\rm IA}$ (see Sect. 5.2) and the redshift offset δ_{z_i} for each tomographic bin i (see Sect. 3). This is a stronger assumption than what is required by the data consistency, since the IA signal, which depends on the galaxy population, is not expected to be the same for the two sub-samples.

Prior distributions for all the free parameters are listed in Table 3. The common nuisance parameters adopt priors from H20, where $A_{\rm IA}$ has a wide flat prior, whereas δ_{z_i} have Gaussian priors with variance determined from a spatial bootstrapping approach during the redshift calibration (see Sect. 3.2 of H20). The six new test parameters in the test model $\mathcal{H}_1$ use wide and uninformative priors. As will be shown in Sect. 7, these prior choices incorporate prior knowledge of redshift uncertainties into the common nuisance parameters and meanwhile allow for a thorough exploration of the test parameters. We stress that the main goal of our test is to evaluate the sufficiency of the KV450 nuisance parameters in capturing residual systematics.

Since we do not rely on the Bayesian evidence to diagnose tensions, our test method is free from the "suspiciousness" problem linked to common model-selection methods (Lemos et al. 2019); in this respect, our test approach is analogous to the second tier of the Bayesian consistency tests proposed by Köhlinger et al. (2019). However, instead of duplicating the cosmological parameters and drawing conclusions based on the posterior distributions of cosmological parameter differences, we focus on the nuisance parameters, especially those linked to the redshift calibration. The other essential difference is that we performed a colour-based split of the source galaxies and re-did measurements and calibrations for the sub-samples, whereas Köhlinger et al. (2019) based their comparison on a split of the measured correlation functions. Therefore, our method is more sensitive to possible inconsistencies within the source samples, whereas their approach is a more global test of residual systematics and the impact on the final cosmological results. In this sense, our test serves as a complementary check of the pipeline robustness to theirs.

7. Results

The main results from our consistency tests are shown in Fig. 7. These are the marginal posterior constraints of the five test parameters $\delta_{z_i,{\rm s}}$ introduced in Sect. 6.2. The five sections in the plot correspond to the five tomographic bins. The two sets of values are from the two sets of cosmological parameters we employed: the KV450 best-fit cosmology (red lines) and the *Planck* best-fit cosmology (black lines). Both sets of results agree with each other, further confirming that our test model is insensitive to the choice of cosmological parameters. As can be seen, all values are consistent with zero within $\sim 1.5\sigma$, indicating that the KV450 calibration pipelines are correcting these sample-related systematics, and introducing more nuisance parameters is unnecessary for the current analysis.

The two tomographic bins with slight non-vanishing differences are the second bin ($\sim 1.2\sigma$) and the third bin ($\sim 1.3\sigma$). Interpreting this level of difference is complex, given the statistical power of the current data. We reiterate that the $\delta_{z_i,{\rm s}}$ parameters we constrained here refer to the shifts of the redshift offsets in the two sub-samples. These are expected to be larger than the mean redshift offsets (δ_{z_i}), given the substantial redshift differences between the two sub-samples and the width of the DIR redshift distributions (see Fig. 2). As seen from Table 3, all $\delta_{z_i,{\rm s}}$ values are smaller than the width of the underlying redshift distributions and are close to zero within the uncertainties. This reflects the overall accuracy of the DIR redshift distributions.

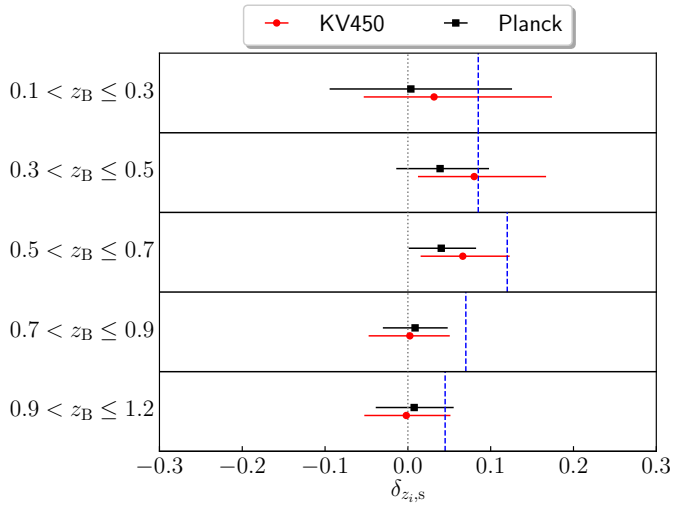

Fig. 7. Constraints on $\delta_{z_i,s}$ per tomographic bin for the $\mathcal{H}_1$ model. Errors shown correspond to the 68% credible intervals from the MCMC run. For comparison, the vertical blue lines show the half of the mean differences between the reconstructed DIR redshift distributions of the two sub-samples (see Fig. 2).

Table 3 lists the posterior results for all free parameters and the best-fit χ^2-values for all models. We do not base our conclusion on the χ^2-test, because the dimensionality is not directly specified by the number of free parameters in a complex Bayesian model (see e.g Handley & Lemos 2019a). Nevertheless, a simple comparison of the best-fit χ^2 values with the number of free parameters taken into account suggests that the test model $\mathcal{H}_1$ is indistinguishable from the control model $\mathcal{H}_0$. This lends some more credit to our previous conclusion on the adequacy of current nuisance parameters in dealing with residual systematics.

Figure 8 presents the contour plot for the test model. An interesting feature we note is the high degeneracy between $A_{\rm IA,s}$ and $\delta_{z_i,s}$ in the low redshift bins (see Fig. 8). This incurs most of the ambiguities in the test parameters. The entanglement between the IA signal and the redshift uncertainties is also noticed in Wright et al. (2020b), where a revised redshift calibration of the KV450 data results in a vanishing IA amplitude. Our finding affirms the difficulty in interpreting the apparent IA signal. We conducted an extreme test where we fixed $\delta_{z_i,s} = 0$ in the test model $\mathcal{H}_1$. It led to a large positive $A_{\rm IA,s}$ value, suggesting $A_{\rm IA,blue} > A_{\rm IA,red}$. This is inconsistent with dedicated IA studies (see Joachimi et al. 2015, for a review), implying that IA parameters can disguise problems with the redshift estimates. Therefore, we should be careful to interpret the IA parameters. To check the impact of the IA parameters in our test model, we ran one more test $\mathcal{T}_1$, in which $A_{\rm IA,s}$ was fixed to zero. This maximises the shifts of the redshift offsets by ignoring the IA difference in the two sub-samples. Even in this conservative estimate, the shifts are $\lesssim 2.1\sigma$ for all redshift bins, with the highest values again seen in the third bin (see Table. 3).

8. Summary and discussion

We presented an internal-consistency test to the KV450 cosmic shear analysis with a colour-based split of source galaxies, resulting in two statistically comparable sub-samples containing noticeably different galaxy populations (see Figs. 1, 2 and 3). We performed the same measurements and calibrations to these two sub-samples and assessed changes in the two-point correlation functions because of known differences in the redshift distributions and the multiplicative biases (see Fig. 5). By fixing cosmological parameters, we examined the internal consistency of the observational nuisance parameters, specifically those for the redshift distributions, using a Bayesian analysis with dedicated test parameters. We observed a degeneracy between the redshift uncertainties and the inferred IA amplitude for low redshift bins, but we found no evidence of internal inconsistency in the KV450 data, verifying that the current strategy of linearly shifting redshift distributions with a common set of nuisance parameters is adequate for capturing residual systematics in the redshift calibration.

The internal-consistency test we proposed is robust against the uncertainties of the background cosmology and cosmic variance. It can be implemented in future cosmic shear surveys before any cosmological inference is made. This weak sensitivity to cosmology is shared with the existing "shear-ratio" test (Jain & Taylor 2003; Schneider 2016; Unruh et al. 2019), which has already been applied to check the accuracy of redshift distributions in current cosmic shear surveys (Heymans et al. 2012b; H20; Giblin et al. 2020). The "shear-ratio" test is a cross-correlation approach based on the galaxy-galaxy lensing signals of two or more source samples at different redshift bins. Therefore, the two tests are sensitive to different systematics, making them complementary.

Although our discussion concentrated on the redshift calibration, we found that the test also relies on our assumptions regarding the IA signals (see Fig. 8). Without a thorough exploration of IA models, our test can already pick up the degeneracy between the IA signals and the redshift uncertainties, which has been implied in previous studies (see Sect. 6.6 of Hildebrandt et al. 2017). Recently, Samuroff et al. (2019) performed an analogous split-based analysis to the DES data. They focus on the IA signal and cosmological parameters and marginalise over observational nuisance parameters. This is different from what we explored here, but connected to our test through the IA signals, which were examined in both tests. They provided better constraints on the IA signals in sub-samples using a variety of IA models. We can perform analogous improvements to our test model to learn more about the IA signals and their correlation to other nuisance parameters in future cosmic shear data.

Acknowledgements. We thank Shahab Joudaki for carefully reading the manuscript and providing useful comments. SSL is supported by NOVA, the Netherlands Research School for Astronomy. KK acknowledges support by the Alexander von Humboldt Foundation. HHo acknowledges support from Vici grant 639.043.512, financed by the Netherlands Organisation for Scientific Research (NWO). HHi is supported by a Heisenberg grant of the Deutsche Forschungsgemeinschaft (Hi 1495/5-1) as well as an ERC Consolidator Grant (No. 770935). This work is based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 100.A-0613, 102.A-0047, 179.A-2004, 177.A-3016, 177.A-3017, 177.A-3018, 298.A-5015, and on data products produced by the KiDS consortium.

Author Contributions: All authors contributed to the development and writing of this paper. The authorship list is given in two groups: the lead authors (SSL, KK, HHo), followed by an alphabetical group of key contributors to both the scientific analysis and the data products.

References

Audren, B., Lesgourgues, J., Benabed, K., & Prunet, S. 2013, J. Cosmology Astropart. Phys., 2013, 001
Bartelmann, M. & Schneider, P. 2001, Phys. Rep., 340, 291
Benítez, N. 2000, ApJ, 536, 571

Table 3. Priors and posterior results for all models.

Parameter	Prior	KV450			Planck		
		$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{T}_1$	$\mathcal{H}_0$	$\mathcal{H}_1$	$\mathcal{T}_1$
$A_{\rm IA}$	$[-6, 6]$	$1.442^{+0.826}_{-0.898}$	$1.049^{+0.818}_{-0.871}$	$0.976^{+0.776}_{-0.804}$	$1.741^{+0.507}_{-0.533}$	$1.358^{+0.463}_{-0.495}$	$1.340^{+0.466}_{-0.476}$
δ_{z_1}	0.000 ± 0.039	$-0.012^{+0.037}_{-0.037}$	$-0.000^{+0.035}_{-0.038}$	$0.001^{+0.035}_{-0.037}$	$-0.037^{+0.028}_{-0.036}$	$-0.008^{+0.036}_{-0.040}$	$-0.005^{+0.038}_{-0.039}$
δ_{z_2}	0.000 ± 0.023	$-0.006^{+0.019}_{-0.023}$	$-0.001^{+0.022}_{-0.021}$	$-0.000^{+0.021}_{-0.022}$	$-0.011^{+0.019}_{-0.019}$	$-0.003^{+0.020}_{-0.022}$	$-0.002^{+0.021}_{-0.019}$
δ_{z_3}	0.000 ± 0.026	$0.009^{+0.023}_{-0.022}$	$0.006^{+0.022}_{-0.023}$	$0.006^{+0.022}_{-0.026}$	$0.020^{+0.020}_{-0.018}$	$0.019^{+0.020}_{-0.021}$	$0.021^{+0.020}_{-0.020}$
δ_{z_4}	0.000 ± 0.012	$-0.002^{+0.012}_{-0.011}$	$-0.001^{+0.012}_{-0.011}$	$-0.002^{+0.012}_{-0.012}$	$0.003^{+0.011}_{-0.012}$	$0.003^{+0.012}_{-0.012}$	$0.003^{+0.012}_{-0.013}$
δ_{z_5}	0.000 ± 0.011	$0.002^{+0.011}_{-0.011}$	$0.003^{+0.012}_{-0.010}$	$0.002^{+0.011}_{-0.011}$	$0.006^{+0.012}_{-0.011}$	$0.005^{+0.011}_{-0.010}$	$0.006^{+0.011}_{-0.011}$
$A_{\rm IA,s}$	$[-6, 6]$	-	$0.571^{+1.178}_{-1.337}$	-	-	$0.536^{+0.793}_{-0.967}$	-
$\delta_{z_1,\rm s}$	$[-0.3, 0.3]$	-	$0.032^{+0.142}_{-0.085}$	$0.079^{+0.076}_{-0.069}$	-	$0.004^{+0.122}_{-0.098}$	$0.072^{+0.057}_{-0.066}$
$\delta_{z_2,\rm s}$	$[-0.3, 0.3]$	-	$0.080^{+0.087}_{-0.068}$	$0.116^{+0.048}_{-0.055}$	-	$0.039^{+0.059}_{-0.053}$	$0.069^{+0.032}_{-0.033}$
$\delta_{z_3,\rm s}$	$[-0.3, 0.3]$	-	$0.066^{+0.057}_{-0.051}$	$0.087^{+0.037}_{-0.041}$	-	$0.040^{+0.042}_{-0.039}$	$0.060^{+0.027}_{-0.030}$
$\delta_{z_4,\rm s}$	$[-0.3, 0.3]$	-	$0.002^{+0.048}_{-0.050}$	$0.014^{+0.044}_{-0.045}$	-	$0.009^{+0.039}_{-0.039}$	$0.019^{+0.037}_{-0.037}$
$\delta_{z_5,\rm s}$	$[-0.3, 0.3]$	-	$-0.002^{+0.053}_{-0.051}$	$0.005^{+0.051}_{-0.050}$	-	$0.008^{+0.048}_{-0.046}$	$0.015^{+0.046}_{-0.046}$
$N_{\rm data}$	-	390	390	390	390	390	390
$N_{\rm para}$	-	6	12	11	6	12	11
χ^2	-	366.8	356.4	356.1	357.5	364.5	364.4

Notes. The first six parameters are common nuisance parameters to account for overall IA amplitude and redshift offsets. The following are six test parameters introduced to account for potential differences of aforementioned parameters between the two sub-samples (see Eq. 10). Priors shown in brackets are top-hat ranges whereas values with errors indicate Gaussian distributions. Results are the mean values of the posterior whereas the χ^2 corresponds to the maximum likelihood. Two sets of results were derived, fixing cosmological parameters to either the KV450 or the *Planck* values (see Table. 2). The test model $\mathcal{H}_1$ contains 12 free parameters. The 'control model' $\mathcal{H}_0$ ignores parameter differences between the two sub-samples and only includes 6 common parameters. The test setting $\mathcal{T}_1$ ignores the difference of IA signals in the two sub-samples.

Blas, D., Lesgourgues, J., & Tram, T. 2011, J. Cosmology Astropart. Phys., 2011, 034
Bridle, S., Balan, S. T., Bethge, M., et al. 2010, MNRAS, 405, 2044
Bridle, S. & King, L. 2007, New Journal of Physics, 9, 444
Brinckmann, T. & Lesgourgues, J. 2018, arXiv e-prints, arXiv:1804.07261
Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125
Coe, D., Benítez, N., Sánchez, S. F., et al. 2006, AJ, 132, 926
Coleman, G. D., Wu, C. C., & Weedman, D. W. 1980, ApJS, 43, 393
de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, A&A, 582, A62
de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., et al. 2017, A&A, 604, A134
Duffy, A. R., Schaye, J., Kay, S. T., & Dalla Vecchia, C. 2008, MNRAS, 390, L64
Edge, A., Sutherland, W., Kuijken, K., et al. 2013, The Messenger, 154, 32
Eisenstein, D. J. & Hu, W. 1998, ApJ, 496, 605
Fenech Conti, I., Herbonnet, R., Hoekstra, H., et al. 2017, MNRAS, 467, 1627
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Fortuna, M. C., Hoekstra, H., Joachimi, B., et al. 2020, arXiv e-prints, arXiv:2003.02700
Giblin, B., Heymans, C., Asgari, M., et al. 2020, arXiv e-prints, arXiv:2007.01845
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Griffith, R. L., Cooper, M. C., Newman, J. A., et al. 2012, ApJS, 200, 9
Handley, W. & Lemos, P. 2019a, Phys. Rev. D, 100, 023512
Handley, W. & Lemos, P. 2019b, Phys. Rev. D, 100, 043504
Heymans, C., Van Waerbeke, L., Bacon, D., et al. 2006, MNRAS, 368, 1323
Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012a, MNRAS, 427, 146
Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012b, MNRAS, 427, 146
Hikage, C., Oguri, M., Hamana, T., et al. 2019, PASJ, 71, 43
Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020, A&A, 633, A69
Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454
Hill, A. R., van der Wel, A., Franx, M., et al. 2019, ApJ, 871, 76
Hirata, C. M. & Seljak, U. 2004, Phys. Rev. D, 70, 063526
Hoekstra, H., Herbonnet, R., Muzzin, A., et al. 2015, MNRAS, 449, 685
Jain, B. & Taylor, A. 2003, Phys. Rev. Lett., 91, 141302
Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338
Joachimi, B., Cacciato, M., Kitching, T. D., et al. 2015, Space Sci. Rev., 193, 1
Joachimi, B., Lin, C. A., Asgari, M., et al. 2020, arXiv e-prints, arXiv:2007.01844
Joachimi, B., Schneider, P., & Eifler, T. 2008, A&A, 477, 43
Kafle, P. R., Robotham, A. S. G., Driver, S. P., et al. 2018, MNRAS, 479, 3746
Kaiser, N. 1992, ApJ, 388, 272
Kaiser, N. 1998, ApJ, 498, 26
Kannawadi, A., Hoekstra, H., Miller, L., et al. 2019, A&A, 624, A92
Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, ApJ, 467, 38
Kitching, T. D., Balan, S. T., Bridle, S., et al. 2012, MNRAS, 423, 3163
Kitching, T. D., Miller, L., Heymans, C. E., van Waerbeke, L., & Heavens, A. F. 2008, MNRAS, 390, 149
Köhlinger, F., Joachimi, B., Asgari, M., et al. 2019, MNRAS, 484, 3126
Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, A&A, 559, A14
Lemos, P., Köhlinger, F., Handley, W., et al. 2019, arXiv e-prints, arXiv:1910.07820
Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218
Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, MNRAS, 390, 118
Limber, D. N. 1953, ApJ, 117, 134
Loverde, M. & Afshordi, N. 2008, Phys. Rev. D, 78, 123506
Mandelbaum, R., Rowe, B., Armstrong, R., et al. 2015, MNRAS, 450, 2963
Massey, R., Heymans, C., Bergé, J., et al. 2007, MNRAS, 376, 13
Mead, A. J., Heymans, C., Lombriser, L., et al. 2016, MNRAS, 459, 1468
Miller, L., Heymans, C., Kitching, T. D., et al. 2013, MNRAS, 429, 2858
Miller, L., Kitching, T. D., Heymans, C., Heavens, A. F., & van Waerbeke, L. 2007, MNRAS, 382, 315
Mo, H., van den Bosch, F. C., & White, S. 2010, Galaxy Formation and Evolution
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563
Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2018, arXiv e-prints, arXiv:1807.06209
Raichoor, A., Mei, S., Erben, T., et al. 2014, ApJ, 797, 102
Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nature Astronomy, 3, 212
Samuroff, S., Blazek, J., Troxel, M. A., et al. 2019, MNRAS, 489, 5453

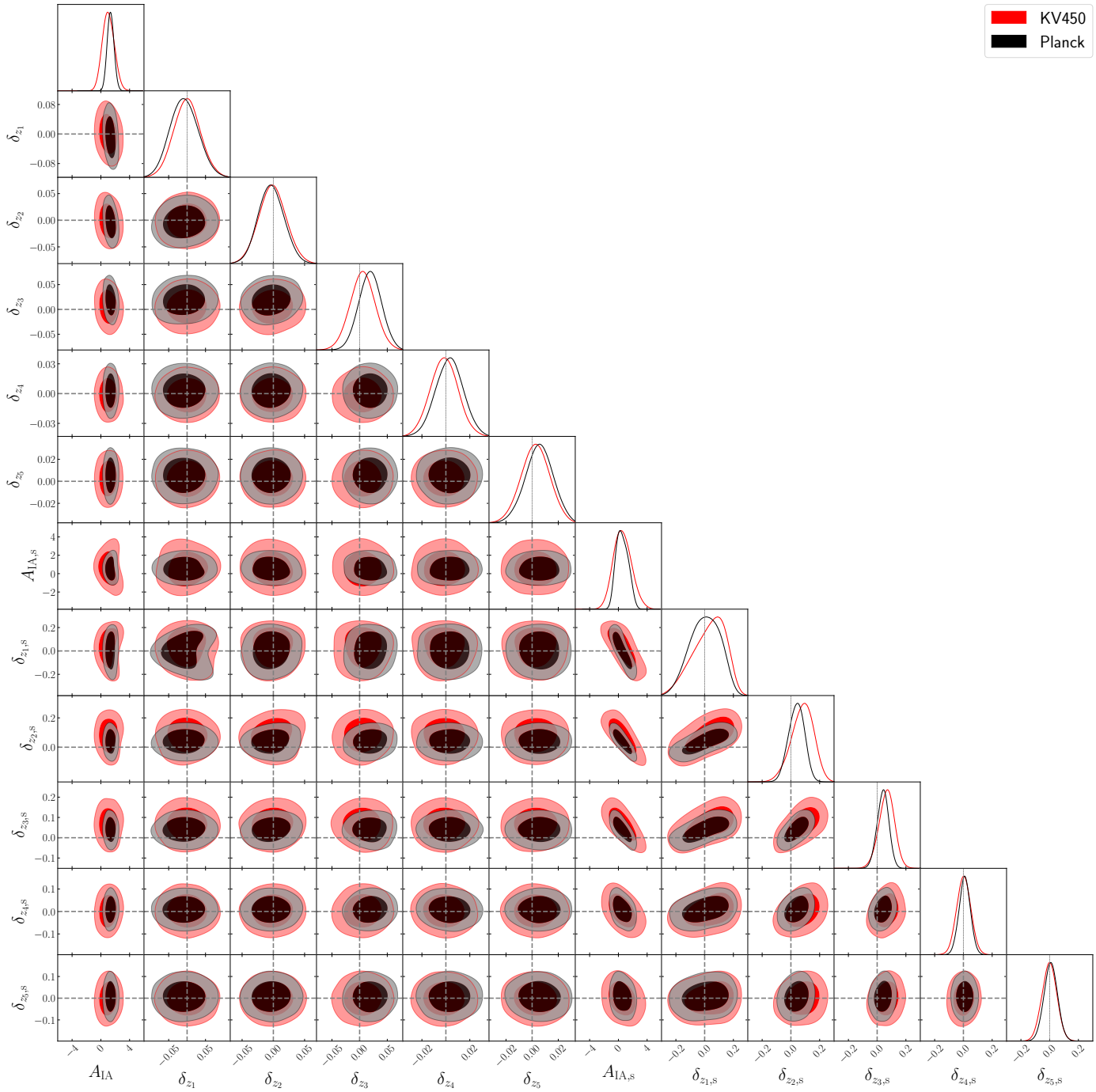

Fig. 8. Contour plots of the 68% and 95% credible regions for all the free parameters in $\mathcal{H}_1$ model. Plotting ranges are the same as the prior ranges. Dashed lines indicate zero values in the ideal case. Two different colours correspond to the two sets of results from KV450 and *Planck* cosmological values. The slight degeneracy between $\delta_{z_i,\rm s}$ in the low redshift bins is an effect from the high degeneracy between $A_{\rm IA,s}$ and $\delta_{z_i,\rm s}$. It vanishes in the test setting $\mathcal{T}_1$, where $A_{\rm IA,s}$ is fixed to zero.

Schneider, P. 2016, A&A, 592, L6
Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, ApJ, 546, 20
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Takada, M. & Hu, W. 2013, Phys. Rev. D, 87, 123504
Takahashi, R., Sato, M., Nishimichi, T., Taruya, A., & Oguri, M. 2012, ApJ, 761, 152
Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, ApJ, 724, 878
Trotta, R. 2008, Contemporary Physics, 49, 71
Troxel, M. A. & Ishak, M. 2015, Phys. Rep., 558, 1
Troxel, M. A., MacCrann, N., Zuntz, J., et al. 2018, Phys. Rev. D, 98, 043528
Unruh, S., Schneider, P., & Hilbert, S. 2019, A&A, 623, A94
Viola, M., Kitching, T. D., & Joachimi, B. 2014, MNRAS, 439, 1909
Wright, A. H., Hildebrandt, H., Kuijken, K., et al. 2019, A&A, 632, A34
Wright, A. H., Hildebrandt, H., van den Busch, J. L., & Heymans, C. 2020a, A&A, 637, A100
Wright, A. H., Hildebrandt, H., van den Busch, J. L., et al. 2020b, arXiv e-prints, arXiv:2005.04207