KiDS+VIKING-450: Consistency tests for cosmic shear tomography with a colour-based split of source galaxies

Shun-Sheng Li\textsuperscript{1}, Konrad Kuijken\textsuperscript{1}, Henk Hoekstra\textsuperscript{1}, Hendrik Hildebrandt\textsuperscript{2}, Benjamin Joachimi\textsuperscript{3}, and Arun Kannawadi\textsuperscript{4}

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

Received XXX, XXXX; accepted YYY, YYYY

ABSTRACT

We perform 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 subsamples, we inspect the characteristics of the shear measurements and the performance of the calibration pipelines. On the modelling side, we examine the observational nuisance parameters, specifically those for the redshift calibration and intrinsic alignments, using a Bayesian analysis with dedicated test parameters. We verify 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

The current standard model of cosmology, dubbed Λ cold dark matter (ΛCDM), is widely accepted given its ability to describe a wide variety of observations spanning early- and late-universe probes. Despite the general agreement, however, some emerging anomalies raise concerns about the correctness of the ΛCDM model and the fidelity of various cosmological probes (see Verde et al. 2019, for a recent review). One of these discrepancies is the mild tension in the amplitude of matter density fluctuations: cosmic microwave background (CMB) constraints favour a more clumped Universe compared to what is preferred by cosmic shear surveys, which yield amplitudes that are 5–10% lower.

Cosmic shear, the coherent distortion of distant galaxies that arises from weak gravitational lensing by large-scale structures, is sensitive to the amplitude of matter density fluctuations, usually quantified by \( \sigma_8 \), and to the mean matter density \( \Omega_m \). Therefore, the main result from a cosmic shear survey is conventionally reported as a derived parameter \( S_8 \equiv \sigma_8 (\Omega_m/0.3)^{0.5} \). Alternatively, CMB measurements infer the local density fluctuations by extrapolating the measured amplitude of temperature fluctuations at recombination, assuming a cosmological model. For 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 in mild tension with results from recent cosmic shear surveys, such as the Dark Energy Survey (DES; Troxel et al. 2018b, \( S_8 = 0.782^{+0.027}_{-0.022} \)), the Hyper Suprime-Cam Subaru Strategic Program (HSC; Hikage et al. 2019, \( S_8 = 0.780^{+0.030}_{-0.033} \)), and the Kilo-Degree Survey (KiDS; Hildebrandt et al. 2020, hereafter H20, \( S_8 = 0.737^{+0.040}_{-0.036} \)).

Although this level of inconsistency is a concern in the era of “precision cosmology”, we have to be careful about any potential systematic effects associated with observations before asserting the failure of the ΛCDM model or the existence of new physics. Given this consideration, performing internal-consistency checks is a standard part of any cosmological probe. Although there is some discussion about the internal consistency of the Planck measurements (Addison et al. 2016; Motloch & Hu 2018, 2020), much of the discussion focuses on the local cosmological probes (e.g. Troxel et al. 2018a; Köhlinger et al. 2019).

A cosmic shear study typically bases its 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 the doubling of 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 systematics may differ.

Source galaxy properties challenge the calibration pipelines mainly in two aspects: the shape measurements and the redshift estimates. First, the shape measurements are sensitive to the dis-

---

\( \sigma_8 \): The standard deviation of linear-theory density fluctuations in a sphere of radius \( 8h^{-1} \) Mpc, where \( H_0 = 100h \) km s\(^{-1}\) Mpc\(^{-1}\).
tributions of galaxy ellipticities, e.g. the lensfit algorithm used in the KiDS survey assigns weights to the measured ellipticities, resulting in a bias toward intermediate ellipticity values (Fenech Conti et al. 2017). The ellipticity distributions are in turn linked to galaxy types, 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). The value of the shear bias is thus related to the underlying galaxy sample. Second, both the accuracy and the precision of a photometric redshift estimate depends on broad spectral features of a galaxy, e.g. the 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 subsamples according to their spectral types and apply the same measurement and calibration pipelines to these two subsamples. This way we explore 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 perform a Bayesian analysis with dedicated test parameters describing relative deviations of the nuisance parameters between the two subsamples. 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 explore a different aspect: we fix cosmological parameters but explore changes in the nuisance parameters instead. We find 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 perform redshift calibration in Sect. 3 and shear bias calibration in Sect. 4. We then measure and model the shear signal in Sect. 5. We introduce the covariance matrix and conduct 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) 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 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 \(\epsilon_1, \epsilon_2\)). 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 lensfit 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).

Throughout we only use sources with valid nine-band photometry (GAAP_Flag_ugriZYJKs==0). This mask reduces the original area by \(~5\%\) and retains \(~13\) million objects. Following H20, we bin source galaxies into five tomographic bins defined as \(0.1 < z_B \leq 0.3, 0.3 < z_B < 0.5, 0.5 < z_B < 0.7, 0.7 < z_B \leq 0.9, 0.9 < z_B \leq 1.2\). Given our purpose of checking systematic effects caused by galaxy properties, we further split the whole sample into two subsamples 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 define our two subsamples 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 starburst types, labelled as blue in this paper). This cut is chosen to ensure similar statistical power in the two subsamples (see Fig. 1). Source properties of these two subsamples are summarised in Table 1.

\(^2\) https://github.com/lshuns/CosmicShearRB

\(^3\) http://kids.strw.leidenuniv.nl/DR3/kv450data.php
Table 1. Source information in the two subsamples.

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

Notes. The effective number density \(n_{\text{eff}}\) 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 are computed from the dispersion of 50 bootstrap samples. The mean and median of the redshift distributions are obtained from the DIR calibration, which is detailed in Sect. 3.

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 subsamples, separately. We follow 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 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 use 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-G15 Deep 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 have 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 (\(agrizYJK\)). 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 adopt the same sample and split it with the same criterion as used for the main KV450 sample to build two representatives of our two subsamples.

The resulting redshift distributions of the two subsamples 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 subsample, where an overall bias toward overestimating photo-z is shown. This may seem counterintuitive at first, but we stress that the red subsample 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 purpose, we are interested in the redshift difference between the two subsamples. 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 subsamples (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 \(\delta_i\) in their model to allow for linear shifts of the redshift distributions \(n_i(z) \rightarrow n_i(z + \delta_i)\) (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 subsamples (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 under-
stand 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 (Hoekstra et al. 2015). Following Heymans et al. (2006), we quantify these residual biases using a linear parameterisation

\[ g_i^{\text{obs}} = (1 + m_i)g_i^{\text{true}} + c_i , \]  

where \( g_i^{\text{obs}} \) and \( g_i^{\text{true}} \) are the observed and the true gravitational shears, respectively, with \( i = 1, 2 \) referring to the two different components. In practice, we find 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-estimate multiplicative biases in the two subsamples 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 find noticeable differences in the galaxy properties for the two subsamples. 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 bias (see also Viola et al. 2014) and therefore an indication of the variance of shear biases in the two subsamples.

Fig. 2. Redshift distributions for the two subsamples, estimated from DIR technique. Shaded regions correspond to photo-\( z \) cuts for the tomographic binning. Mean and median differences are calculated as \( \delta z_{\text{mean/median}} = z_{\text{mean/median,blue}} - z_{\text{mean/median,red}} \).
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 subsamples in Fig. 4, compared with the results from the whole sample. The five sections from top to bottom correspond to the five tomographic bins from lower to higher redshifts. We see 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 (see Sect. 5). We thus assume 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 \( \delta_c \) is introduced into the model to account for a potential offset of the empirically determined \( \epsilon_1 \) values. The result from forward-modelling suggests that \( \delta_c \) is very close to 0 (see Table 2). Third, a position-dependent additive bias pattern in the \( \epsilon_1 \) 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 follow this strategy for the additive bias calibration. We correct the \( \epsilon \)-term per tomographic bin and per patch using the same empirical approach mentioned above. We also include the 2D \( \epsilon \)-term pattern in our test models. But we abandon the two nuisance parameters \( \delta_c \) and \( A_c \) from our model, as they do not have a significant impact on the fit.
The joint data vector (\(\mathbf{x}\)) are chosen to mitigate baryon feedback on small scales and first seven bins for \(\xi\). Mysterically spaced bins within the interval \([0, 1]\) are used. The measurements contains (7 \(\pm\) 3 \(\xi\), \(\pm\) 3 \(\xi\), \(\pm\) 3 \(\xi\), \(\pm\) 3 \(\xi\)). Using the Kaiser-Limber approximation (Limber 1953; Kaiser 1992, 1998; Loverde & Afshordi 2008), \(\xi(q)\) is in turn related to the physical matter power spectrum \(P_m\), via

\[
P_m(q) = \int_0^\infty \frac{d\chi}{\chi} \frac{q_i(\chi)q_j(\chi)}{[f_k(\chi)]^2} P_m(\ell + 1/2) \frac{f_k(\chi')}{f_k(\chi)},
\]

where \(\chi\) and \(f_k(\chi)\) are the comoving radial distance and the comoving angular distance, respectively. The upper limit of the integral \(\chi_H\) is the comoving horizon distance. The lensing efficiency \(q_i(\chi)\) for tomographic bin \(i\) is defined as

\[
q_i(\chi) = \frac{3H_0^2 \Omega_m f_k(\chi)}{2c^2} \int_0^\infty \frac{d\chi'}{\chi'} n_i(\chi') f_k(\chi' - \chi) \frac{f_k(\chi')}{f_k(\chi)},
\]

which depends on the redshift distribution of galaxies \(n_i(\chi) d\chi = n_i(z) dz\) along with other cosmological parameters. Therefore, different redshift distributions will cause a difference in shear signal between the two subsamples.

We calculate 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 assume a \(\Lambda\) CDM 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 (\(\Omega_{CDM}\) and \(\Omega_b\)), the amplitude and the index of the scalar power spectrum (\(\sigma_8\)), the scaled Hubble parameter \((h)\), and the amplitude of the halo mass-concentration relation (\(B\)).

For our purpose of consistency tests, it is unnecessary to explore this whole cosmological parameter space, which is the same for the two subsamples. Therefore, we fix 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 test 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 “non-linear linear” IA model into the measured shear signal (Hirata & Seljak 2004; Bridle & King 2007):
as $\xi_{\text{in},\text{in}}$ ("intrinsic-intrinsic" term between the intrinsic ellipticities of nearby galaxies) and $\xi_{\text{II},\text{in}}$ ("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

$$P_{\text{II}}^\text{in}(\ell) = \int_0^\infty d\chi F_\chi^2(z) \frac{n_\chi n_\chi(\chi)}{[f_\chi(\chi)]^2} P_\chi\left(\ell + 1/2, f_\chi(\chi), \chi\right),$$

(7)

$$P_{\text{II}}^\text{II}(\ell) = \int_0^\infty d\chi F_\chi(z) q_\chi n_\chi(\chi)+q_\chi n_\chi(\chi) P_\chi\left(\ell + 1/2, f_\chi(\chi), \chi\right),$$

(8)

where

$$F(z) = -A_{\text{IA}} \rho_{\text{crit},0} \chi(z).$$

(9)

The normalisation constant is $C = 5 \times 10^{-14} h^{-1} M_\odot^{-1} \text{Mpc}^3$, $\rho_{\text{crit},0}$ is the critical density today, and the linear growth factor $D_\chi(z)$ is normalised to unity today. Following H20, we ignore the redshift and luminosity dependence of IA and leave one nuisance parameter $A_{\text{IA}}$ 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 fix all the model parameters and use the redshift distributions estimated in Sect. 3 to predict the joint data vector of the two subsamples. 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 subsamples. The other feature worth to note is the similarity between the two predictions from the two different sets of cosmological parameters. This implies that our test models are 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
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 |
|-----------|-------|--------|
| $\Omega_{\text{CDM}} h^2$ | 0.058 | 0.120 |
| $\Omega_b h^2$ | 0.022 | 0.022 |
| ln($10^{10} A_s$) | 4.697 | 3.045 |
| $n_s$ | 1.128 | 0.966 |
| $h$ | 0.780 | 0.673 |
| $B$ | 2.189 | - |
| $A_{\text{IA}}$ | 0.494 | - |
| $\delta_i \times 10^5$ | 2.576 | - |
| $A_c$ | 1.143 | - |
| $\delta_{i1}$ | -0.006 | - Bin 1 offset |
| $\delta_{i2}$ | 0.001 | - Bin 2 offset |
| $\delta_{c1}$ | 0.026 | - Bin 3 offset |
| $\delta_{c2}$ | -0.002 | - Bin 4 offset |
| $\delta_{\tau}$ | 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 Planck 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 build an analytical covariance matrix to account for all the correlations between measurements (Sect. 6.1). We then perform a Bayesian analysis with dedicated test parameters to quantify the potential discrepancy between measurements from the two subsamples (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 test models.

The modelling pipeline detailed below is publicly available6. It is a modified version of the MultiNest algorithm (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).

6.1. Covariance matrix

We estimate 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 ($n_g$) and the weighted ellipticity dispersion ($\sigma_{\epsilon,g}$). The second, non-Gaussian term is calculated using the formalism from Takada & Hsu (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 adopt a systematic uncertainty $\sigma_{\text{sys}} = 0.02$ for the multiplicative biases as estimated by K19 and used in H20 and Wright et al. (2020b) and propagate it into the covariance matrix through $C_{ij}^{\text{sys}} = 4\xi_i^2 \Sigma_j^2 + C_{ij}$, where $\xi_i^2$ is the joint data vector predicted using the KV450 best-fit values and the DIR redshift distributions (see Sect. 3). We ignore 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 subsamples and across the two subsamples. The importance of the potential correlations between (two) parts of a split was already highlighted in Köhlinger et al. (2019), but here we confirm it more directly. By including the full covariance matrix into our consistency tests, we naturally take 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 subsamples, we constructed the covariance matrix of $\Delta \xi$ as $C_{\Delta \xi} = C_{\text{blue}} + C_{\text{red}} - 2C_{\text{cross}}$, and compared it to the covariance matrices of the single data vectors ($\xi_{\text{blue}}$ or $\xi_{\text{red}}$). We found that the non-

---

6 https://github.com/lishuns/montepython_KV450
Gaussian contributions are significantly suppressed in \( C_{A} \) with an overall reduction of \( \leq 75\% \) compared to \( C_{\text{blue}} \). 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 \( C_{A} \). Therefore, we verify 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 subsamples. For this purpose, we fix all the cosmological parameters, which in principle share the same values in the two subsamples. Fixing these parameters simplifies the likelihood function and avoids unnecessary exploration of the high-dimensional parameter space. To account for any potential residual effects from an “incorrect” choice of cosmological parameters, we run two setups with cosmological parameters from the KV450 cosmic shear results and from the Planck CMB results (see Table 2).

We consider two main hypotheses for the choice of nuisance parameters. In the null hypothesis \( \mathcal{H}_{0} \), we assume that a common set of nuisance parameters is satisfactory for the residual systematics in the two subsamples. It includes six free nuisance parameters: the amplitude of the IA signal \( A_{\text{IA}} \) (see Sect. 5.2) and the redshift offset \( \delta_{s} \) 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 subsamples. In the alternative hypothesis \( \mathcal{H}_{1} \), we assume that it requires more nuisance parameters to capture differences between the two subsamples. It thus adds six more test parameters besides the common nuisance parameters from the \( \mathcal{H}_{0} \) hypothesis: a shift in IA amplitude \( A_{\text{IA}} \), and shifts in redshift offsets \( \delta_{s, \text{DIR}} \). We implement them in the two subsamples as

\[
X_{\text{blue/red}} = X \pm X_{s},
\]

where \( X \) represents the \( A_{\text{IA}} \) or \( \delta_{s} \) parameters, whereas \( X_{s} \) denotes corresponding test parameters. The plus sign is applied to the blue subsample, and the minus sign is for the red subsamples. 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 \( \delta_{s, \text{DIR}} \) imply residual systematics that cannot be adequately captured by the common nuisance parameters. We therefore base our result mainly on the posterior distributions of these test parameters.

Prior distributions for all the free parameters are listed in Table 3. The common nuisance parameters adopt priors from \( H_{20} \), where \( A_{\text{IA}} \) has a wide flat prior, whereas \( \delta_{s} \) have Gaussian priors with variance determined from a spatial bootstrapping approach during the redshift calibration (see Sect. 3.2 of \( H_{20} \)). The six new test parameters in hypothesis \( \mathcal{H}_{1} \) 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 perform a colour-based split of the source galaxies and redo measurements and calibrations for the subsamples, whereas Köhlinger et al. (2019) base 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 \( \delta_{s, \text{DIR}} \) 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 employ: 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 \( \delta_{s, \text{DIR}} \) parameters we constrained here refer to the shifts of the redshift offsets in the two subsamples. These are expected to be larger than the mean redshift offsets (\( \delta_{s} \)), given the substantial redshift differences between the two subsamples and the width of the DIR redshift distributions (see Fig. 2). As seen from Table 3, all
δ_{z,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. Table 3 lists the posterior results for all free parameters and the best-fit $\chi^2$-values for all hypotheses. We do not base our conclusion on the $\chi^2$-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 $\chi^2$ values with the number of free parameters taken into account suggests that the two main hypotheses are indistinguishable from each other. 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 $\mathcal{H}_1$ hypothesis. An interesting feature we note is the high degeneracy between $A_{IA,s}$ and $\delta_{z,s}$ 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 fix $\delta_{z,s} = 0$ in the alternative hypothesis $\mathcal{H}_1$. It leads to a large positive $A_{IA,s}$ value, suggesting $A_{IA,blue} > A_{IA,red}$. 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 run one more test $T_1$, in which $A_{IA,s}$ is fixed to zero. This maximises the shifts of the redshift offsets by ignoring the IA difference in the two subsamples. Even in this conservative estimate, the shifts are $\leq 2.1\sigma$ for all redshift bins, with the highest values again seen in the third bin (see Table 3).

8. Summary and Discussion

We have presented an internal-consistency test to the KV450 cosmic shear analysis with a colour-based split of source galaxies, resulting in two statistically comparable subsamples containing noticeably different galaxy populations (see Figs. 1, 2 and 3). We perform the same measurements and calibrations to these two subsamples and assess 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 examine the internal consistency of the observational nuisance parameters, specifically those for the redshift distributions, using a Bayesian analysis with dedicated test parameters. We observe a degeneracy between the redshift uncertainties and the inferred IA amplitude for low redshift bins, but we find 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 propose 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 has concentrated on the redshift calibration, we find 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) have 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 explore here, but connected to our test through the IA signals, which are examined in both tests. They provide better constraints on the IA signals in subsamples 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). HHo 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

Addison, G. E., Huang, Y., Watts, D. J., et al. 2016, ApJ, 818, 132
Table 3. Priors and posterior results for the test models.

| Parameter | Prior | KV450 | Planck |
|-----------|-------|-------|--------|
| \(A_{\Lambda}\) | \([-6, 6]\) | \([0.004 \pm 0.826, 0.049 \pm 0.181]\) | \([0.976 \pm 0.776, 1.741 \pm 0.507]\) | \([1.358 \pm 0.463, 1.340 \pm 0.466]\) |
| \(\delta_{c1}\) | \(0.000 \pm 0.039\) | \([-0.012 \pm 0.037, -0.000 \pm 0.035]\) | \([0.001 \pm 0.035, -0.037 \pm 0.028]\) | \([-0.008 \pm 0.036, -0.040 \pm 0.005]\) |
| \(\delta_{c2}\) | \(0.000 \pm 0.023\) | \([-0.006 \pm 0.019, -0.001 \pm 0.022]\) | \([-0.000 \pm 0.021, -0.011 \pm 0.019]\) | \([-0.003 \pm 0.020, -0.022 \pm 0.021]\) |
| \(\delta_{c5}\) | \(0.000 \pm 0.026\) | \([0.009 \pm 0.023, 0.006 \pm 0.022]\) | \([0.006 \pm 0.022, 0.020 \pm 0.020]\) | \([0.019 \pm 0.020, 0.021 \pm 0.020]\) |
| \(\delta_{c6}\) | \(-0.000 \pm 0.012\) | \([-0.022 \pm 0.012, -0.001 \pm 0.012]\) | \([-0.002 \pm 0.012, 0.003 \pm 0.011]\) | \([0.003 \pm 0.012, -0.012 \pm 0.013]\) |
| \(\delta_{c7}\) | \(0.000 \pm 0.011\) | \([0.002 \pm 0.011, 0.003 \pm 0.012]\) | \([-0.002 \pm 0.011, 0.006 \pm 0.012]\) | \([-0.005 \pm 0.011, 0.006 \pm 0.011]\) |

\(N_{\text{data}}\) | - | 390 | 390 | 390 | 390 | 390 |
| \(N_{\text{para}}\) | 6 | 12 | 11 | 6 | 12 | 11 |
| \(\chi^2\) | - | 366.8 | 365.4 | 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 subsamples (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 \(\chi^2\) corresponds to the maximum likelihood. Two sets of results are derived, fixing cosmological parameters to either the KV450 or the Planck values (see Table 2). The alternative hypothesis \(H_1\) is the fiducial test model with 12 free parameters. The null hypothesis \(H_0\) ignores parameter differences between the two subsamples and only includes 6 common parameters. The test setting \(T_1\) ignores the difference of IA signals in the two subsamples.

S.-S. Li et al.: KiDS+VIKING-450: Consistency tests for cosmic shear tomography
\[ \delta z_1, s \]

\[ \delta z_2, s \]

\[ \delta z_3, s \]

\[ \delta z_4, s \]

\[ \delta z_5, s \]

\[ KV450 \]

\[ Planck \]

Fig. 8. Contour plots of the 68% and 95% credible regions for all the free parameters in $H_1$ hypothesis. 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.