Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **484,** 3126–3153 (2019)
Advance Access publication 2019 January 16

doi:10.1093/mnras/stz132

A Bayesian quantification of consistency in correlated data sets

Fabian Köhlinger,[1⋆] Benjamin Joachimi,[2⋆] Marika Asgari,[3] Massimo Viola,[4] Shahab Joudaki[5] and Tilman Tröster[3]

[1]*Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan*
[2]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[3]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[4]*Leiden Observatory, Leiden University, PO Box 9513, Leiden, NL-2300 RA, the Netherlands*
[5]*Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*

Accepted 2019 January 9. Received 2018 December 23; in original form 2018 September 5

ABSTRACT

We present three tiers of Bayesian consistency tests for the general case of *correlated* data sets. Building on duplicates of the model parameters assigned to each data set, these tests range from Bayesian evidence ratios as a global summary statistic, to posterior distributions of model parameter differences, to consistency tests in the data domain derived from posterior predictive distributions. For each test, we motivate meaningful threshold criteria for the internal consistency of data sets. Without loss of generality we focus on mutually exclusive, correlated subsets of the same data set in this work. As an application, we revisit the consistency analysis of the two-point weak-lensing shear correlation functions measured from KiDS-450 data. We split this data set according to large versus small angular scales, tomographic redshift bin combinations, and estimator type. We do not find any evidence for significant internal tension in the KiDS-450 data, with significances below $3\,\sigma$ in all cases. Software and data used in this analysis can be found at http://kids.strw.leidenuniv.nl/sciencedata.php.

Key words: gravitational lensing: weak – methods: data analysis – statistical – cosmology: cosmological parameters – observations – large-scale structure of Universe.

1 INTRODUCTION

The key objective in any kind of data analysis is to find a model which describes the data the best and with the fewest assumptions following Occam's razor. Once such a model has been established there are typically two more questions arising:

(i) Is the data set self-consistent (under the given model)?
(ii) Is the data set consistent with another data set (under the given model)?

Without loss of generality, we will address both questions in the context of parameter inference from cosmological probes.

The current cosmological concordance model is rooted in General Relativity including at least a cosmological constant (Λ) and the yet-to-be directly detected cold dark matter (CDM). The success and acceptance of the ΛCDM model lies in its ability to explain a wide range of cosmological observables such as the fluctuation spectrum of the cosmic microwave background (CMB) radiation, distance measurements with supernovae of type Ia, the clustering of galaxies and the gravitational lensing of the cosmic large-scale structure with just a handful of parameters.

⋆ E-mail: fabian.koehlinger@ipmu.jp (FK); b.joachimi@ucl.ac.uk (BJ)

However, there currently exist discrepancies between parameters inferred from different cosmological probes leading us back to the two questions posed at the beginning. The statistically most significant example for such a discrepancy is that the Hubble constant inferred from CMB measurements of the *Planck* satellite (Planck Collaboration XIII 2016, VI 2018) disagree with the value derived from local measurements (Riess et al. 2016, 2018) by $3.4\,\sigma$–$3.6\,\sigma$. Moreover, the cosmological parameters controlling the scaling of the weak-lensing signal amplitude measured from the cosmic large-scale structure show further inconsistencies with *Planck* CMB measurements (Planck Collaboration XIII 2016) ranging from 1.7σ for the Dark Energy Survey (DES, DES Collaboration 2017) to $2.3\,\sigma$–$3.2\,\sigma$ for the Kilo-Degree Survey (KiDS, Hildebrandt et al. 2017; Köhlinger et al. 2017).

In the current era of precision cosmology driven by increasingly larger surveys with lower and lower statistical errors, it is thus paramount to identify the sources for these discrepancies as either arising from residual systematics in the cosmological probe(s), insufficient modelling of observables and nuisances, or new physics. Hence, methods and tests to assess the self-consistency of a data set and to assess the consistency between different cosmological probes become evermore important. For the latter case various approaches have been proposed in the literature mainly assessing the consistency of *Planck* with respect to other statistically *independent*

cosmological probes (see for example Charnock, Battye & Moss 2017, and references therein for a summary of such methods; see also Lin & Ishak 2017a,b; Adhikari & Huterer 2018; Raveri & Hu 2018 for recent developments).

The self-consistency of the *Planck* measurements, for example the consistency of the low and high multipole measurements was assessed in Planck Collaboration LI (2017) based on a cross-validation approach and found discrepancies at $\leqslant 2.2\,\sigma$. Moreover, Efstathiou & Lemos (2018) recently presented a consistency check of the weak-lensing correlation function measurements from 450 sq deg of KiDS imaging data (KiDS-450 henceforth, Hildebrandt et al. 2017) based on the same cross-validation approach. They found significant tension at $\gtrsim 3\,\sigma$ when for example the correlation function measurements were split into mutually exclusive subsets containing different combinations of the redshift distributions of the source galaxies. A caveat of the cross-validation approach is that the data set under consideration is typically split into two independent parts and using only one of the two, the other part is predicted. This approach necessarily neglects all intrinsic correlations between the (two) parts of the split.

In contrast to this approach in this paper, we develop three tiers of consistency tests for the most general case of *correlated* data sets that take all correlations between the data sets fully into account. This is achieved by basing the tests on duplicated model parameters for each data set. The test statistics then include Bayesian evidence ratios as a global summary statistic, posterior probability density functions (PDFs or PDF henceforth) of model parameter differences and consistency tests in the data domain derived from posterior predictive distributions. With these tests at hand we will revise the self-consistency of the KiDS-450 analysis split into various mutually exclusive subsets serving at the same time as a test case and example of a *correlated* data set. This gives us ample opportunity to contrast and critically discuss the advantages and disadvantages of each approach to consistency.

The paper is structured as follows: in Section 2, we present the methodology for the three tiers of consistency tests. In Section 3, we provide a pedagogical guide to our approach through an analytically tractable toy model, further supported by an extensive sensitivity analysis of the proposed consistency tests in a realistic setting in Appendix A. The KiDS-450 cosmic shear data and its likelihood are briefly described in Section 4. Finally, we apply our consistency tests to the KiDS-450 data and also compare this approach to the cross-validation approach of Efstathiou & Lemos (2018) in Section 5 before concluding in Section 6.

2 METHODOLOGY

In the following, we will develop three tiers of consistency tests. First, we use the Bayes factor as a global summary statistic for consistency/tension. The Bayes factor alone does not provide us with a diagnostic of where discrepancies may be present in the data, and it may fail to flag issues that only affect a subset of the full data set. Therefore, we add two additional diagnostics: one in parameter space based on duplicates of the model parameters, and one in the vector space spanned by the original data, which we refer to as the data domain.

2.1 First tier: Bayes factor

The guiding principle for the consistency test in this section is the question: 'How much more probable is it that the full data set was generated from the same model system than if each individual (sub)data set were generated from an independent set of model parameters?'. Addressing that quantitatively by making use of Bayesian evidences as proposed by Marshall, Rajguru & Slosar (2006) yields a conservative test to quantify the consistency between measurements of cosmological parameters from uncorrelated data sets.

The Bayesian evidence, $\mathcal{Z}$, is simply the normalization factor occurring in the calculation of a posterior PDF (and often neglected when only parameter inference is of interest). For an N-dimensional data vector $\boldsymbol{d}$, parameters $\boldsymbol{p}$, and a model (or hypothesis) H, it gives the average of the likelihood times the prior PDF/probability over the M-dimensional parameter space:

$$\mathcal{Z} = \Pr(\boldsymbol{d}\,|\,\mathrm{H}) = \int \mathrm{d}^M \boldsymbol{p}\,\Pr(\boldsymbol{d}\,|\,\boldsymbol{p},\mathrm{H})\,\Pr(\boldsymbol{p}\,|\,\mathrm{H}), \tag{1}$$

where $\Pr(\boldsymbol{d}\,|\,\boldsymbol{p},\mathrm{H})$ is the likelihood of producing the data given the parameters of the model H and $\Pr(\boldsymbol{p}\,|\,\mathrm{H})$ is the prior for a given set of parameters of the model H. In that sense, the Bayesian evidence has Occam's razor built in: a model that requires more parameters has a lower evidence than a more compact model, unless the more complex model describes the data significantly better. Therefore, comparing evidence ratios presents a meaningful way of selecting one model (or hypothesis) over the other. That also implies that the evidence increases with increasing goodness of fit for a given model and decreases for more complicated models for a given goodness of fit, where 'complicated' may imply additional parameters and/or a larger prior volume. In the case of comparing and quantifying dis-/concordance of data sets, we are less interested in comparing different (nested) models to each other but rather in comparing the probabilities of the two statements:

(i) H_0 : 'there exists one common set of parameters that describes all data sets' and

(ii) H_1 : 'there exist more than one set of parameters that each describe one data set'.

Hence, we write down their probability ratio as:

$$\frac{\Pr(\mathrm{H}_0\,|\,\boldsymbol{d})}{\Pr(\mathrm{H}_1\,|\,\boldsymbol{d})} = \frac{\Pr(\boldsymbol{d}\,|\,\mathrm{H}_0)}{\Pr(\boldsymbol{d}\,|\,\mathrm{H}_1)}\,\frac{\Pr(\mathrm{H}_0)}{\Pr(\mathrm{H}_1)}. \tag{2}$$

In the case that there is no a priori reason to prefer one model over the other, which we assume throughout the remainder of the text, the probability ratio reduces to comparing the evidence ratio, also referred to as the Bayes factor,

$$\mathrm{R}_{01} = \frac{\Pr(\boldsymbol{d}\,|\,\mathrm{H}_0)}{\Pr(\boldsymbol{d}\,|\,\mathrm{H}_1)} = \frac{\Pr(\boldsymbol{d}\,|\,\mathrm{H}_0)}{\prod_i \Pr(\boldsymbol{d}_\mathrm{i}\,|\,\mathrm{H}_1)}, \tag{3}$$

where the full data vector has been split into subsets, $\boldsymbol{d}^\tau = \{\boldsymbol{d}^\tau_\mathrm{a}, \boldsymbol{d}^\tau_\mathrm{b}, \ldots\}$. It is important to realize that the right-hand-side of equation (3) holds only if the data sets $\boldsymbol{d}_\mathrm{i}$ are *independent* of each other. This is not necessarily the case if we want to quantify the consistency of measurements from the *same* data set (e.g. different splits of the data, and different estimators). In that case we indeed need to evaluate the first expression of equation (3), i.e. we require the cross-covariance between the data sets $\boldsymbol{d}_\mathrm{i}$ and need to keep i independent sets of parameters while evaluating the joint likelihood.

Generally, a Bayes factor of $R_{01} < 1$ in this test is an indicator for tension between the data sets. For a more detailed interpretation of the evidence ratio we use its common logarithm and the quantitative scale by Jeffreys (1961).

In this work, we consider only a single split of a data set, i.e. $\boldsymbol{d}^\tau = \{\boldsymbol{d}^\tau_\mathrm{a}, \boldsymbol{d}^\tau_\mathrm{b}\}$. Each of the two subsets, $\boldsymbol{d}_\mathrm{a}$ and $\boldsymbol{d}_\mathrm{b}$, gets assigned

its own copy of the full parameter set, $\boldsymbol{p}$.[1] Then, for case H_1, the joint posterior and the evidence of the duplicate parameter sets is inferred from the full data set in order to calculate the Bayes factor, R_{01}, with respect to the evidence of the fiducial single parameter set of case H_0, i.e.

$$R_{01} = \frac{\Pr(\boldsymbol{d} \mid H_0)}{\Pr(\{\boldsymbol{d}_a^\tau, \boldsymbol{d}_b^\tau\} \mid H_1)}. \quad (4)$$

In practice, we evaluate the nominator and denominator of equation (4) while estimating the best-fitting parameters, $\boldsymbol{p}$, by means of a customized likelihood evaluation code (see Section 4.1).

2.2 Second tier: differences of parameter duplicates

The joint posterior of the duplicated parameter sets used to evaluate the denominator of equation (4) also provides us with another valuable diagnostic: we can now derive posterior PDFs of the differences between the two instances of the same parameter and identify cases where these difference distributions are inconsistent with zero and hence reveal potential biases in the posteriors. This constitutes the second tier of consistency tests.

We quantify tension for this tier by determining how unlikely it is to end up in a region with lower posterior probability density than the origin, since the origin marks the point of perfect agreement between the subsets in the space of parameter differences. In this work, we restrict ourselves to the marginal posterior of the three most interesting and constraining parameters. The approach is implemented as follows: we apply kernel density estimation with a Gaussian kernel (Scott 1992) to the Markov Chain Monte Carlo (MCMC) sample to obtain a functional form of the posterior. The posterior density at the origin is evaluated and the fraction of MCMC samples with lower density values calculated. The lower this fraction the more extreme is the location of the origin relative to the region of high posterior density, thus indicating tension with the expectation of zero parameter difference for a given split.

We propose to cast the tension estimate into the popular formulation of '$m\,\sigma$', which is linked to the probability mass c_m within the range $[-m\,\sigma, m\,\sigma]$ of a one-dimensional Gaussian, i.e. $c_m = 0.683$ for $m = 1$, $c_m = 0.954$ for $m = 2$, and etc. The aforementioned fraction is then identified with $1 - c_m$.

2.3 Third tier: predictive distributions

The evidence ratio test outlined above is compressing a lot of information into a single number and only answers the question of whether the data sets are in tension. Moreover, if the model is a bad description of the data in the first place, the Bayes factor will not flag this either. If tension is detected, using this test gives us only very limited information about its origin.

Therefore, we propose as a third complementary diagnostic tool, predictions of the data vector from previously inferred posteriors of the model parameters. Traditionally, this is achieved via the posterior predictive distribution (PPD henceforth). It is the sampling distribution for new data $\hat{\boldsymbol{d}}$ given the existing data $\boldsymbol{d}$ under the model

[1]Note that this is the most conservative choice. It may be useful to also consider only duplicating a subset of the parameter set, e.g. creating copies of the cosmological parameters while keeping a single set of nuisance parameters. We leave the study of such applications to future work.

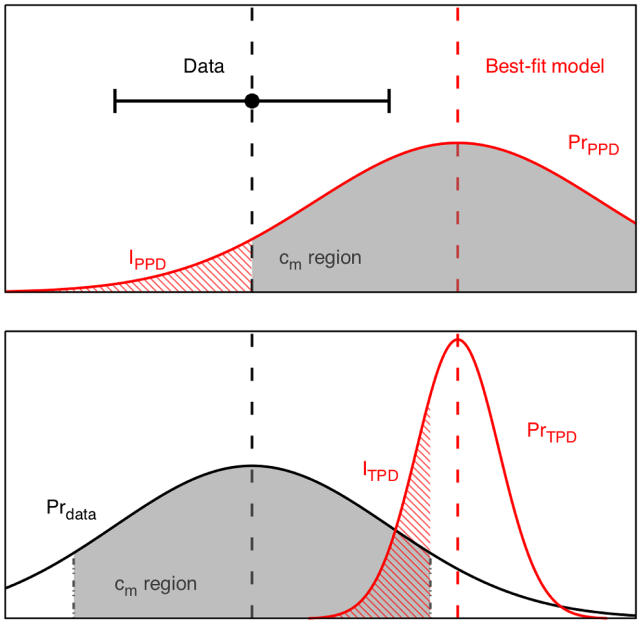

Figure 1. Sketch illustrating the definition of significance criteria for tension between data and model predictions. Top: PPD case. The c_m region is defined as the support of the PPD where its density is higher than the density at the position of the data. The hatched area, given by $I_{PPD} = 1 - c_m$, is used to derive the tension significance. Bottom: TPD case. The fraction of the TPD probability mass lying in the support of the c_m region of the data distribution (I_{TPD}, hatched area) is calculated. If this fraction drops below $1 - c_m$, the distributions are in tension by $m\,\sigma$.

H_α, i.e. one averages the likelihood of the new data over the posterior of the parameters $\boldsymbol{p}$:

$$\Pr(\hat{\boldsymbol{d}} \mid \boldsymbol{d}, H_\alpha) = \int \mathrm{d}^M \boldsymbol{p}_\alpha \Pr(\hat{\boldsymbol{d}} \mid \boldsymbol{p}_\alpha, H_\alpha) \Pr(\boldsymbol{p}_\alpha \mid \boldsymbol{d}, H_\alpha). \quad (5)$$

To test for tension in the data, one should check if the actual data vector $\boldsymbol{d}$ is incompatible with being a sample drawn from the PPD. For that several (summary) statistics are possible (e.g. Gelman et al. 2013). For example, the properties of the PPD can be characterized via an ensemble of synthetic data vectors $\hat{\boldsymbol{d}}$ drawn from it. Feeney et al. (2018) quantify tension by calculating the ratio of the PPD probability density at the data and the PPD mode. Instead, we employ again the $m\,\sigma$ formalism: one determines the probability mass in the region where the probability density of the PPD is higher than the value at the data. The complement, I_{PPD}, of this region is then identified with $1 - c_m$, as is illustrated in the top panel of Fig. 1.

The PPD will in general inherit a non-Gaussian shape from the posterior and therefore not be analytic and typically be available in the form of a Monte Carlo sample. Its dimension is that of the original data vector and thus of order 100 or more in many cases of interest. This makes a consistency test via the PPD, as outlined above, impractical. Instead, we introduce a novel approach that we will demonstrate to have very similar performance, and that keeps consistency tests in high-dimensional spaces tractable if the likelihood is Gaussian.

To this end, we define a predicted data vector, $\boldsymbol{d}_{pre}$, which is uniquely determined as a model prediction for a given set of parameter values, $\boldsymbol{p}$. Thus, we can use $\boldsymbol{d}_{pre}$ to rephrase the PPD

of equation (5) as

$$\Pr(\hat{\boldsymbol{d}} \mid \boldsymbol{d},\ \mathrm{H}_\alpha) = \int \mathrm{d}^N \boldsymbol{d}_{\rm pre}\, \Pr(\hat{\boldsymbol{d}} \mid \boldsymbol{d}_{\rm pre}, \mathrm{H}_\alpha)\, \Pr(\boldsymbol{d}_{\rm pre} \mid \boldsymbol{d}, \mathrm{H}_\alpha). \tag{6}$$

If the likelihood is Gaussian, the first probability on the right-hand side is given by

$$\ln \Pr(\hat{\boldsymbol{d}} \mid \boldsymbol{d}_{\rm pre}, \mathrm{H}_\alpha) = -\frac{1}{2}(\hat{\boldsymbol{d}} - \boldsymbol{d}_{\rm pre})^{\intercal}\, \mathbf{C}^{-1} (\hat{\boldsymbol{d}} - \boldsymbol{d}_{\rm pre}) + \text{const.}\,, \tag{7}$$

where $\mathbf{C}$ is the covariance matrix of the data. The second probability in the integral of equation (6) is simply a translation of the posterior into the data domain, i.e. in practice, one calculates a model prediction for every Monte Carlo sample in parameter space. We shall refer to this probability as the translated posterior distribution, or TPD. It quantifies the spread of possible model predictions given the uncertainty on the model parameters.

The TPD and the 'data distribution' of equation (7) are special cases of the PPD. The former results if there is zero measurement error, i.e. $\Pr(\hat{\boldsymbol{d}} \mid \boldsymbol{d}_{\rm pre}, \mathrm{H}_\alpha) \to \delta_{\rm D}(\hat{\boldsymbol{d}} - \boldsymbol{d}_{\rm pre})$ in equation (6), where $\delta_{\rm D}$ is the Dirac delta distribution. The latter results if the model is perfect, i.e. it has zero uncertainty and recovers each data point exactly, and thus $\Pr(\boldsymbol{d}_{\rm pre} \mid \boldsymbol{d}, \mathrm{H}_\alpha) \to \delta_{\rm D}(\boldsymbol{d}_{\rm pre} - \boldsymbol{d})$. We propose to use a comparison between the TPD and the data distribution in equation (7) as a consistency check. If the predictions of the actual model and a perfect model agree within the uncertainties of the inferred model and the measurement error, we have a consistent data set (under that model).

The quantitative analysis now amounts to a comparison of two distributions of which one (equation 7) is widely assumed to be a multivariate Gaussian and therefore known analytically. This is readily extended to high dimensions, as detailed below. We follow Charnock et al. (2017) in quantifying tension between the distributions by integrating the TPD over the iso-contours of a given significance level of the data distribution:

$$\mathrm{I}_{\rm TPD} = \int_{V_{\rm data}} \mathrm{d}^N \boldsymbol{x}\, \Pr_{\rm TPD}(\boldsymbol{x}) \text{ with} \tag{8}$$

$$V_{\rm data} = \left\{ \boldsymbol{x} \,\middle|\, \int_{V_{\rm data}} \mathrm{d}^N \boldsymbol{x}\, \Pr_{\rm data}(\boldsymbol{x}) = c_{\rm m} \right\}, \tag{9}$$

where $\Pr_{\rm TPD}$ denotes the TPD, and $\Pr_{\rm data}$ the data distribution. The integral in equation (9) is understood to be over the subvolume(s) of the data domain in which $\Pr_{\rm data}(\boldsymbol{x})$ attains its highest values. This definition of tension reproduces the intuitive expectation in the case of a low-dimensional Gaussian distribution in that it measures the shift of the mean of the Gaussian in units of its standard deviation (see Fig. 4 below for an illustration).

Charnock et al. (2017) propose to call the two distributions to be in tension by $m\,\sigma$ if $\mathrm{I}_{\rm TPD} = 0$ beyond the $m\,\sigma$-level. In contrast to that, we increase this threshold from zero to $\mathrm{I}_{\rm TPD} = 1 - c_{\rm m}$. This has the major advantage that the definition of tension becomes independent of the number of samples drawn in practice from the TPD. This significance criterion is illustrated in the bottom panel of Fig. 1.

In practice, we evaluate the integral $\mathrm{I}_{\rm TPD}$ in equation (8) by calculating the quantity

$$\chi^2 = (\boldsymbol{d}_{\rm data} - \boldsymbol{d}_{\rm TPD})^{\intercal}\, \mathbf{C}^{-1}\, (\boldsymbol{d}_{\rm data} - \boldsymbol{d}_{\rm TPD}), \tag{10}$$

between the data vector, $\boldsymbol{d}_{\rm data}$, and each TPD vector, $\boldsymbol{d}_{\rm TPD}$, derived from the typically of order 10^4 MCMC samples. We then read off limits, $\chi^2_{\rm lim}$, from the chi-squared distribution with N degrees of freedom which correspond to the $m\,\sigma$ levels of the N-dimensional (Gaussian) data distribution. The values of $\chi^2_{\rm lim}$ define a surface within which a fraction $c_{\rm m}$ of the probability mass of the data distribution is contained. Note that we will use the same approach on the real data as it is assumed to follow a Gaussian likelihood.

In practice, we obtain the TPD by translating every Monte Carlo sample in parameter space (e.g. as readily available from the calculations for the first two tiers of consistency checks described in Sections 2.1 and 2.2) back into the data domain. For the significance calculation we then determine the fraction of TPD samples for which the value of χ^2 according to equation (10) is below $\chi^2_{\rm lim}$. We calculate this integral for $m\sigma$ levels in the range $0 \leq m \leq 10$ with a step size of $\delta m = 0.01$.

Predictive distributions are often used in a cross-validation approach, i.e. a model posterior is inferred from one subset of the full data vector, and a predicted data vector derived for the other subset. We will perform analyses following this philosophy in Section 5.3, but for the majority of this paper will use the full data vector $\boldsymbol{d}$ to infer the posterior and then predict replicas of $\boldsymbol{d}$ via the predictive distributions. This has the advantage of keeping the analysis symmetric, while in cross-validation mode, the choice of subset used for the model inference may lead to different conclusions.

Since we have two types of posterior, one for the standard, 'joint' inference and one for the duplicate parameter set, the 'split' analysis, we can also construct two corresponding types of predictive distributions. A tension between the joint and split TPDs suggests an unaccounted for systematic effect that affects one subset significantly more than the other. This comparison constitutes our second TPD-based consistency estimator. In practice, we calculate the difference

$$\Delta^{\rm TPD}_{\rm j,s_{a/b}} = \boldsymbol{d}^{\rm TPD}_{\rm joint} - \boldsymbol{d}^{\rm TPD}_{\rm split_{a/b}} \tag{11}$$

and assign a significance for the tension between the joint and split TPDs by fitting $\Delta^{\rm TPD}_{\rm j,s_{a/b}}$ to zero and quantifying its deviation from zero by comparison to a chi-squared distribution. For this, we also need to calculate an inverse covariance matrix of the $\Delta^{\rm TPD}_{\rm j,s_{a/b}}$ estimator, which is non-trivial due to the expected strong correlations between the joint and split TPDs. The details of this calculation can be found in Appendix B.

We emphasize once more that the estimator defined in equation (11) quantifies an unaccounted systematic effect affecting one subset more than the other. In contrast but quite complementary to that, the first TPD-based estimator defined in equations (8) and (10) quantifies a tension between the split TPDs and the data distribution. It is thus indicative of trends in the data (systematic or physical) not captured by the model that affect both subsets in a similar manner, and therefore cannot be absorbed through the flexibility of the duplicated parameters.

Further intuition on the workings of the proposed consistency tests can be gained from the sensitivity analysis of mock weak-lensing data provided in Appendix A. In the following Section 3, we provide an analytically tractable worked example of all three tiers. Readers interested in real data should skip ahead to Section 4 and following.

3 A WORKED EXAMPLE

To guide the intuition of the reader for the three tiers of consistency tests introduced in the previous sections, we present in this section analytically tractable toy models. For example, in Fig. 2 we consider

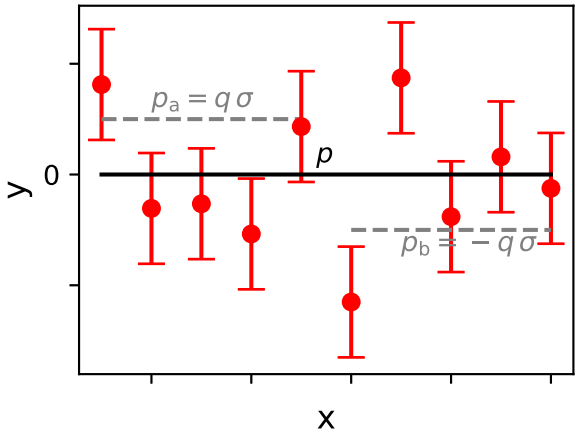

Figure 2. Sketch of a simple toy model consisting of N independent data points (red) drawn from a normal distribution with width σ. The data can be modelled with a constant line with a free amplitude p (black line). If the data set is split into two subsets (of equal size), we allow each subset to be modelled with shifted amplitudes $p_a = q\,\sigma$ and $p_b = -q\,\sigma$, respectively.

N independent data points, $\boldsymbol{d}$, drawn from a Gaussian with variance σ^2 which can be described by a simple model: a constant line with a free amplitude p as the single parameter. The corresponding likelihood function can be written as

$$\mathrm{Pr}_0(\boldsymbol{d}|p) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}|\boldsymbol{d}-\boldsymbol{m}|^2\sigma^{-2}\right), \tag{12}$$

where the model is given by $\boldsymbol{m} = (\overbrace{p,\ldots,p}^{N})^\tau$. Moreover, we assume a Gaussian prior on the amplitude p with width Δ and without loss of generality centred on zero:

$$\mathrm{Pr}_0(p) = \frac{1}{\sqrt{2\pi}\Delta} \exp\left(-\frac{p^2}{2\Delta^2}\right). \tag{13}$$

When splitting the data into two subsets, i.e. $\boldsymbol{d}^\tau = \{\boldsymbol{d}_a^\tau, \boldsymbol{d}_b^\tau\}$, the corresponding likelihood function and prior become:

$$\mathrm{Pr}_1(\boldsymbol{d}|[p_a, p_b]) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}|\boldsymbol{d}-\boldsymbol{m}_s|^2\sigma^{-2}\right), \tag{14}$$

with $\boldsymbol{m}_s = (\underbrace{p_a,\ldots,p_a}_{S}, \underbrace{p_b,\ldots,p_b}_{N-S})^\tau$ for S and $N-S$ elements in $\boldsymbol{d}_a$ and $\boldsymbol{d}_b$, respectively. The corresponding prior becomes then:

$$\mathrm{Pr}_1(p_a, p_b) = \frac{1}{2\pi\Delta^2} \exp\left(-\frac{p_a^2+p_b^2}{2\Delta^2}\right). \tag{15}$$

Based on these definitions, we can calculate analytically the statistics of the three tiers of consistency checks as introduced in the previous Sections 2.1–2.3. For the first tier, the Bayes factor, we write down the evidences for the 'joint' data case with subscript '0' and the 'split' data case with subscript '1':

$$\begin{aligned}\mathcal{Z}_0 &= \int \mathrm{d}p\, \mathrm{Pr}_0(\boldsymbol{d}|p)\,\mathrm{Pr}_0(p)\\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-(N-1)}(N\Delta^2+\sigma^2)^{-\frac{1}{2}}\\ &\quad\times\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{N}d_i^2 - \frac{\Delta^2}{\sigma^2(N\Delta^2+\sigma^2)}\right.\right.\\ &\quad\left.\left.\times\left(\sum_{i=1}^{N}d_i\right)^2\right]\right\} \text{ and}\end{aligned} \tag{16}$$

$$\mathcal{Z}_1 = \int \mathrm{d}p_a \int \mathrm{d}p_b\, \mathrm{Pr}_1(\boldsymbol{d}|[p_a, p_b])\,\mathrm{Pr}_1(p_a, p_b) \tag{17}$$

$$= \int \mathrm{d}p_a\, \mathrm{Pr}_1(\boldsymbol{d}_a|p_a)\,\mathrm{Pr}_1(p_a) \int \mathrm{d}p_b\, \mathrm{Pr}_1(\boldsymbol{d}_b|p_b)\,\mathrm{Pr}_1(p_b) \tag{18}$$

$$\begin{aligned}&= (2\pi)^{-\frac{N}{2}}\sigma^{-(S-1)}(S\Delta^2+\sigma^2)^{-\frac{1}{2}}\sigma^{-(N-S-1)}\\ &\quad\times[(N-S)\Delta^2+\sigma^2]^{-\frac{1}{2}}\\ &\quad\times\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{N}d_i^2 - \frac{\Delta^2}{\sigma^2(S\Delta^2+\sigma^2)}\left(\sum_{i=1}^{S}d_i\right)^2\right.\right.\\ &\quad\left.\left.-\frac{\Delta^2}{\sigma^2[(N-S)\Delta^2+\sigma^2]}\left(\sum_{i=S+1}^{N}d_i\right)^2\right]\right\}.\end{aligned} \tag{19}$$

The Bayes factor, $R_{01} = \mathcal{Z}_0/\mathcal{Z}_1$, then becomes

$$\begin{aligned}R_{01} &= \sqrt{\frac{\sigma(S\Delta^2+\sigma^2)[(N-S)\Delta^2+\sigma^2]}{(N\Delta^2+\sigma^2)}}\\ &\quad\times\exp\left\{-\frac{\Delta^2}{2\sigma^2}\left[\frac{\left(\sum_{i=1}^{N}d_i\right)^2}{N\Delta^2+\sigma^2} + \frac{\left(\sum_{i=1}^{S}d_i\right)^2}{S\Delta^2+\sigma^2}\right.\right.\\ &\quad\left.\left.+\frac{\left(\sum_{i=S+1}^{N}d_i\right)^2}{(N-S)\Delta^2+\sigma^2}\right]\right\}.\end{aligned} \tag{21}$$

In order to keep the equations for this tier and all others more concise and tractable, we consider now the following specific example case: the data are split into two equally sized samples (i.e. $S = N-S = N/2$) with

$$\langle d_i\rangle = \left\{\begin{array}{ll} p_a = q\sigma\ , & i \le S \\ p_b = -q\sigma\ , & S < i \le N \end{array}\right\}, \tag{22}$$

i.e. we allow the model to be shifted by $\pm q$ units of the standard deviation σ around the truth at zero (see also Fig. 2). Moreover, we assume that the width of the prior is much larger than the standard deviation of the data, i.e. $\Delta/\sigma \gg 1$. Then, we can calculate the expectation value of the (natural logarithm of the) Bayes factor as given in equation (21) as a function of the parameters q, N, Δ, and σ:

$$\langle \ln R_{01}\rangle \approx \ln\left(\frac{\Delta}{\sigma}\frac{\sqrt{N}}{2}\right) - \frac{1+q^2}{2}. \tag{23}$$

We note that the first term on the right-hand side of this equation depends explicitly on the width of the prior, Δ, and we compare different prior widths in Fig. 3(a).

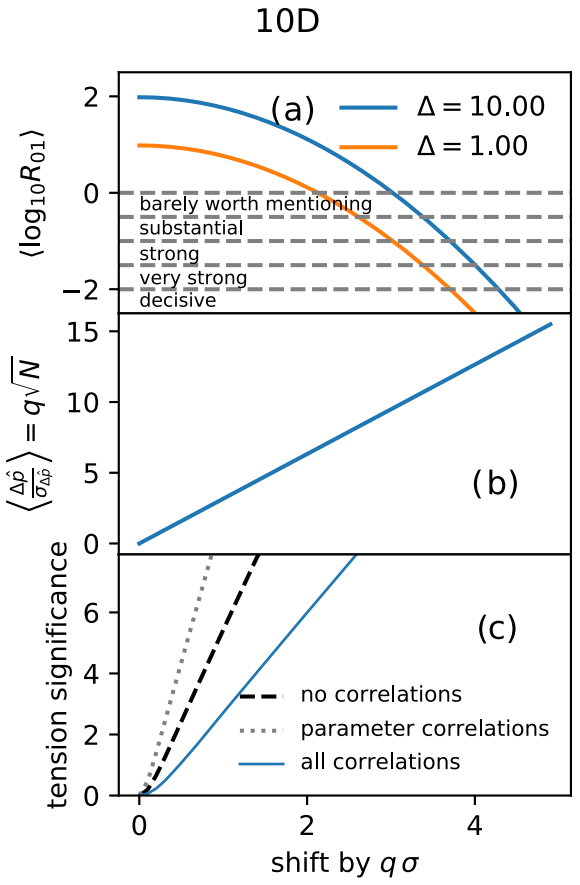

Figure 3. Using the toy model set-up depicted in Fig. 2, i.e. $S = N - S = N/2 = 5$, $\sigma = 0.1$, and hence $\Delta/\sigma \gg 1$ for $\Delta = \{1., 10.\}$, we derive analytically tractable results for the three tiers of consistency tests as functions of the model shift parameter q: (a) the Bayes factor (equation 23). Note that this estimator is the only one strongly depending on the prior width, Δ. We interpret the Bayes factor here in terms of Jeffreys' scale and the statements should be read as 'barely worth mentioning', 'substantial', etc. evidence for H_1: 'there exist two separate parameter sets that each describe one subset of the data'; (b) the relative error of the parameter difference PDF (equation 34); (c) significances for the TPD-based consistency estimator (derived from equation 51). To highlight the impact of a proper propagation of all correlations, we compare the fiducial case of including 'all correlations' (i.e. data subsets and parameters; solid blue line) to the naïve case of 'no correlations' (dashed black line) and 'parameter correlations' only (dotted grey line).

For the second tier, we derive an expression for the differences between the posterior PDFs of the duplicate parameters and the error on it. First, we use Bayes' theorem and equations (12) and (13) to calculate the following proportionality for the posterior:

$$\Pr(p|\boldsymbol{d}) \propto \exp\left[-\frac{1}{2}\left(-\frac{|\boldsymbol{d}-\boldsymbol{m}|^2}{\sigma^2}+\frac{p^2}{\Delta^2}\right)\right] \tag{24}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_i d_i^2 + p^2\left(\frac{N}{\sigma^2}+\frac{1}{\Delta^2}\right) - \frac{2p}{\sigma^2}\sum_i d_i\right]\right\}. \tag{25}$$

Similarly, we find for the posterior PDF of the split sample containing two copies of the parameters, p_a and p_b:

$$\Pr([p_a, p_b]|\boldsymbol{d}) \propto \Pr(p_a|\boldsymbol{d})\Pr(p_b|\boldsymbol{d}) \tag{26}$$

$$= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{S}(d_i-p_a)^2+\frac{p_a^2}{\Delta^2} + \frac{1}{\sigma^2}\sum_{i=S+1}^{N}(d_i-p_b)^2+\frac{p_b^2}{\Delta^2}\right]\right\}, \tag{27}$$

Introducing now the new variables $\bar{p} = (p_a + p_b)/2$ and $\Delta p \equiv p_b - p_a$ lets us rewrite $p_{a/b} = \bar{p} \pm \Delta p/2$ and hence equation (26) becomes:

$$\Pr(\bar{p}, \Delta p|\boldsymbol{d}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{N} d_i^2 + \left(\bar{p}^2+\frac{\Delta p^2}{4}\right)\left(\frac{N}{\sigma^2}+\frac{1}{\Delta^2}\right) - \frac{2\bar{p}}{\sigma^2}\sum_{i=1}^{N} d_i + \frac{\Delta p}{\sigma^2}\left(\sum_{i=1}^{S} d_i - \sum_{i=S+1}^{N} d_i\right)\right]\right\} \tag{28}$$

Marginalizing over $\bar{p}$, we obtain the posterior of the difference in the split parameter,

$$\Pr(\Delta p|\boldsymbol{d}) = \int \mathrm{d}\bar{p}\,\Pr(\bar{p}, \Delta p|\boldsymbol{d}) \tag{29}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{4}\left(\frac{N}{\sigma^2}+\frac{1}{\Delta^2}\right)\right]\left[\Delta p - \left(\sum_{i=S+1}^{N} d_i - \sum_{i=1}^{S} d_i\right)\frac{2}{\sigma^2}\frac{1}{\left(\frac{N}{\sigma^2}+\frac{1}{\Delta^2}\right)}\right]^2\right\} \tag{30}$$

$$\equiv \exp\left[-\frac{1}{2}\left(\frac{\Delta p - \Delta\hat{p}}{\sigma_{\Delta\hat{p}}}\right)^2\right]. \tag{31}$$

Comparing the exponents of equations (30) and (31), we find the following expressions for the mean and variance:

$$\Delta\hat{p} = \frac{2}{N}\left(\sum_{i=S+1}^{N} d_i - \sum_{i=1}^{S} d_i\right)\left(1+\frac{\sigma^2}{N\Delta^2}\right)^{-1} \text{ and} \tag{32}$$

$$\sigma^2_{\Delta\hat{p}} = \frac{4\sigma^2}{N}\left(1+\frac{\sigma^2}{N\Delta^2}\right)^{-1}. \tag{33}$$

Assuming now again the previous toy model case, i.e. $\Delta/\sigma \gg 1$, $S = N - S = N/2$ and $p_a = q\sigma$ and $p_b = -q\sigma$, we can evaluate the expectation value for the relative error of the parameter differences, i.e.

$$\left\langle\left|\frac{\hat{\Delta p}}{\sigma_{\Delta p}}\right|\right\rangle = q\sqrt{N} \tag{34}$$

since $\left(1+\frac{\sigma^2}{N\Delta^2}\right)^{-1} \approx 1$ for $\Delta/\sigma \gg 1$ and $\langle\sum_{i=S+1}^{N} d_i\rangle = -(N-S)q\sigma$ and $\langle\sum_{i=1}^{S} d_i\rangle = Sq\sigma$. We show this estimator as a function of q in Fig. 3, where we also compare it to the estimators of the other tiers.

Finally, we derive an analytic expression for the tension estimator in the third tier of consistency tests (cf. equation 11) in this toy model set-up. From the model vector for the joint sample, $\boldsymbol{m} = (\underbrace{p, \ldots, p}_{N})^\intercal$, and the one for the split sample, $\boldsymbol{m}_s = (\underbrace{p_a, \ldots, p_a}_{S}, \underbrace{p_b, \ldots p_b}_{N-S})^\intercal$, we can define the difference model vector

as $\Delta\boldsymbol{m} = \boldsymbol{m}_\mathrm{j} - \boldsymbol{m}_\mathrm{s}$. It is then straightforward to write down an expression for the χ^2 based on which we will finally assign significances for tension:

$$\chi^2 = \sum_{i=1}^{N} \frac{\Delta m_\mathrm{i}^2}{\sigma^2_{\Delta m}}. \tag{35}$$

Assuming now again our simplified toy model set-up, $\Delta\boldsymbol{m}$ becomes:

$$\Delta m_\mathrm{i} = \left\{ \begin{array}{ll} p - p_\mathrm{a}\ , & i \leq S \\ p - p_\mathrm{b}\ , & S < i \leq N \end{array} \right\} = \left\{ \begin{array}{ll} -q & \sigma \\ q & \sigma \end{array} \right\}. \tag{36}$$

If we assumed no correlation between the two TPDs, then

$$\sigma^2_{\Delta\mathrm{m}} = \left\{ \begin{array}{ll} \sigma_p^2 + \sigma_{p_\mathrm{a}}^2\ , & i \leq S \\ \sigma_p^2 + \sigma_{p_\mathrm{b}}^2\ , & S < i \leq N \end{array} \right\} \approx \left\{ \begin{array}{l} \frac{\sigma^2}{N} + \frac{\sigma^2}{S} \\ \frac{\sigma^2}{N} + \frac{\sigma^2}{N-S} \end{array} \right\} = 3\frac{\sigma^2}{N}, \tag{37}$$

using $\Delta/\sigma \gg 1$ to arrive at the second equality and $S = N - S = N/2$ to arrive at the rightmost equality. This would yield:

$$\chi^2 = \frac{1}{3} N^2 q^2. \tag{38}$$

However, this expression for the χ^2 is overly simplistic since for the calculation of $\sigma^2_{\Delta\mathrm{m}}$ we do need to take into account the correlations between the parameter sets and finally also between the predicted model vectors of the subsamples. We start with the former by employing a Fisher matrix approach similar to what is done in the real data case (see Appendix B).

First we write down a combined parameter vector $\boldsymbol{p} = (p, p_\mathrm{a}, p_\mathrm{b})^\tau$ and, labelling its components in that order with 1, 2, and 3, we can define the Fisher matrix as:

$$(\mathbf{F})_{\mu\nu} = \sum_{i=1}^{N} \frac{\partial m_\mathrm{i}}{\partial p_\mu}\, \sigma^{-2}\, \frac{\partial m_\mathrm{i}}{\partial p_\nu} \tag{39}$$

$$= \sigma^{-2} \begin{pmatrix} N & S & N-S \\ S & S & 0 \\ N-S & 0 & N-S \end{pmatrix}. \tag{40}$$

Evaluating this expression now for $S = N - S = N/2$ yields

$$\mathbf{F} = \frac{N}{\sigma^2} \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}. \tag{41}$$

We immediately realize that $\det \mathbf{F} = 0$, i.e. joint and split parameter sets are fully correlated, but $\mathbf{F}^{-1}$ is needed for the propagation of the parameter correlations. Hence, we diagonalize $\mathbf{F}$ and use a pseudo-inverse to define the correlation matrix:

$$\mathbf{C} \equiv \mathbf{F}^{+} = \frac{\sigma^2}{N} \mathbf{V} \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{V}^\tau = \frac{\sigma^2}{9N} \begin{pmatrix} 4 & 2 & 2 \\ 2 & 10 & 8 \\ 2 & -5 & 10 \end{pmatrix}, \tag{42}$$

with

$$\mathbf{V} = \begin{pmatrix} \sqrt{\frac{2}{3}} & 0 & -\sqrt{\frac{1}{3}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{3}} \\ \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{3}} \end{pmatrix}. \tag{43}$$

Then, we can write:

$$\sigma^2_{\Delta\mathrm{m}} = \left\{ \begin{array}{ll} C_{11} + C_{22} - 2\,C_{12}\ , & i \leq S \\ C_{11} + C_{33} - 2\,C_{13}\ , & S < i \leq N \end{array} \right\} = \frac{10}{9}\frac{\sigma^2}{N}. \tag{44}$$

Plugging this expression now into equation (35) yields

$$\chi^2 = \frac{9}{10} N^2\, q^2. \tag{45}$$

This approach, however, still neglects correlations between the TPD data vectors and to account for that we need to generalize equation (35) to:

$$\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta m_\mathrm{i} [\mathbf{Cov}^{-1}(\Delta\boldsymbol{m})]_{\mathrm{i,j}} \Delta m_\mathrm{j}. \tag{46}$$

The covariance elements that belong to $\Delta\boldsymbol{m}$ within a subset can be adopted from equation (44), while elements across the split are determined from equation (42) as follows:

$$\mathbf{Cov}[\Delta\boldsymbol{m}(\leq S); \Delta\boldsymbol{m}(> S)] = \mathbf{Cov}(p, p) + \mathbf{Cov}(p_\mathrm{a}, p_\mathrm{b}) - \mathbf{Cov}(p, p_\mathrm{a}) - \mathbf{Cov}(p, p_\mathrm{b}) \tag{47}$$

$$= -\frac{8}{9}\frac{\sigma^2}{N}. \tag{48}$$

From that expression, we derive that

$$\mathbf{Cov}(\Delta\boldsymbol{m}) = \frac{\sigma^2}{9N} \begin{pmatrix} 10 & \dots & 10 & -8 & \dots & -8 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 10 & \dots & 10 & -8 & \dots & -8 \\ -8 & \dots & -8 & 10 & \dots & 10 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -8 & \dots & -8 & 10 & \dots & 10 \end{pmatrix}. \tag{49}$$

Using now again $S = N - S = N/2$, we can calculate the pseudo-inverse of that matrix as:

$$\mathbf{Cov}^{+}(\Delta\boldsymbol{m}) = \frac{2}{N\sigma^2} \begin{pmatrix} 5 & \dots & 5 & 4 & \dots & 4 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 5 & \dots & 5 & 4 & \dots & 4 \\ 4 & \dots & 4 & 5 & \dots & 5 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 4 & \dots & 4 & 5 & \dots & 5 \end{pmatrix} \tag{50}$$

Evaluating then equation (46) with that expression for the inverse covariance matrix, we finally find:

$$\chi^2 = N\, q^2. \tag{51}$$

We can then directly assign significances to the χ^2 values following a standard procedure when fitting to zero. The rank of the covariance in equation (49) is 2, which is hence the number of degrees of freedom that should be used to evaluate the goodness-of-fit with equation (51). Analogously, we can expect in a realistic scenario that the degrees of freedom will be of order twice the number of model parameters. In practice, we determine the number of significant eigenvalues of the covariance via principal component analysis (PCA, Appendix B). We show the significances derived from equations (38), (45), and (51) in Fig. 3(c) to demonstrate the effect of correlations with respect to the significances derived from the naïve equation (38): accounting for the correlations introduced by correlated parameter sets (dotted grey line) increases the significances for tension with respect to the naïve case (dashed black line). However, accounting for both parameter correlations

and the correlated data subsamples dilutes the sensitivity of this estimator significantly (solid blue line). In the other panels of Fig. 3, we compare this estimator also to the other two tiers for the same toy model set-up, i.e. equations (23) and (34) in particular. This shows that the sensitivity of the Bayes factor is impaired with respect to the other estimators due to its explicit dependence on the prior width, Δ. Altering it increases or decreases the significance of the Bayes factor while it leaves the other estimators unaffected.

Lastly, we apply the TPD-based goodness-of-fit estimator (equations 8–10) to a more complex toy case that allows us to additionally assess the sensitivity of our significance tests, as well as the impact of noise and correlations. For this we set-up, an N-dimensional mock data vector, $\boldsymbol{d}_{\rm fid}$, drawn from a multivariate Gaussian distribution centred on zero and with an $N \times N$ covariance matrix, $\mathbf{C}$, with entries

$$C_{\rm ij} = r^{|i-j|} s^2, \tag{52}$$

where $0 \leq r < 1$. For $r = 0$, this yields independent data with variance s^2, while $r > 0$ introduces non-trivial correlations. Mock TPDs are created by drawing samples of N-dimensional data vectors $\boldsymbol{d}_{\rm TPD}$ from a multivariate Gaussian distribution centred on zero and with covariance

$$C_{\rm ij}^{\rm TPD} = r^{|i-j|} t^2, \tag{53}$$

where we will choose $t < s$ to reflect that the TPD is typically much more compact than the data distribution. In order to test the way of quantifying tension with the TPDs, we create perturbed data vectors (based on the fiducial realization) by adding to the first $0 \leq Q \leq N$ entries of the vector a constant q, i.e. the mean of the data distribution is given by $\boldsymbol{\mu}^\tau = \{q, \ldots, q, 0, \ldots, 0\}$.

To mimic the process of creating the TPD distributions, we draw by default 1000 samples from $\mathcal{N}\left(\boldsymbol{d}_{\rm TPD}; 0, \mathbf{C}^{\rm TPD}\right)$, i.e. $\boldsymbol{d}_{\rm TPD}$ is Gaussian with mean 0 and covariance $\mathbf{C}^{\rm TPD}$. We then determine the fraction of TPD samples for which the value of χ^2 according to equation (10) is below $\chi^2_{\rm lim}$ as an approximation to calculating the integral $\mathrm{I}_{\rm TPD}$ in equation (8). We calculate this integral for $m\sigma$ levels in the range $0 \leq m \leq 10$ with a step size of $\delta m = 0.01$.

If we additionally restrict ourselves to the case of no correlations ($r = 0$), one can analytically calculate the expected level of significance as follows:

$$\mathrm{I}_{\rm TPD} = \int_{V_N(ms)} \mathrm{d}^N x \, \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{C}^{\rm TPD}\right) \tag{54}$$

$$= \int_{-ms}^{ms} \mathrm{d}x_1 \, \mathcal{N}\left(x_1; \sqrt{Q}q, t^2\right) \int_{V_{N-1}(ms)} \mathrm{d}^{N-1} x \prod_{i=2}^{N} \mathcal{N}\left(x_{\rm i}; 0, t^2\right) \tag{55}$$

$$= \int_{-ms}^{ms} \mathrm{d}x_1 \, \mathcal{N}\left(x_1; \sqrt{Q}q, t^2\right) \left[1 - \frac{\Gamma\left(\frac{N-1}{2}, \frac{m^2 s^2 - x_1^2}{2t^2}\right)}{\Gamma\left(\frac{N-1}{2}, 0\right)}\right]. \tag{56}$$

Here, $V_N(ms)$ denotes the volume of an N-dimensional sphere of radius ms, which defines the support over which the TPD distribution is integrated. We have used the (upper) incomplete Gamma function,

$$\Gamma(a, x) = \int_x^\infty \mathrm{d}y \; y^{a-1} \, \mathrm{e}^{-y}. \tag{57}$$

Note that, purely for notational convenience, we have shifted the TPD distribution by $\boldsymbol{\mu}$, not the data distribution, in equation (54). To arrive at the second equality, we have assumed without loss of generality that the shift vector is aligned with the x_1 axis. We have also used that $|\boldsymbol{\mu}| = \sqrt{Q}q$ in our model. Equation (54) holds for $N \geq 2$; in the 1D case (cf. Fig. 1) the term in square brackets is replaced by unity.

Our definition of tension is intuitive in that in one dimension it corresponds to the shift of the data point μ with respect to the TPD in units of its standard deviation, s. We refer to this as the naive tension criterion. By design, this holds exactly for $t \to 0$, whereas the finite size of the TPD reduces the tension mildly (see Fig. 4a). For comparison, we also consider the definition of tension employed by Efstathiou & Lemos (2018) who calculated the relative deviation from the expected value of their equivalent of equation (10) (see Section 5.3 for a more detailed discussion). In our toy model their significance criterion reads

$$m_{\rm EL18} = \frac{1}{\sqrt{2N}} \left(\frac{\mu^2}{s^2} - N\right), \tag{58}$$

which implies a quadratic dependence on the relative shift of the data vector and hence a stricter notion of tension, with the curves in Fig. 4 rising more sharply than our choice of criterion.

Within the limits of our toy model and no correlations, one can extend the naive tension definition to higher dimensions by using the root mean square of all relative data point shifts, $m_{\rm naive} = \mu/s/\sqrt{N}$. Fig. 4(b) shows the results for 10 dimensions, with the significance of our tension significance criterion lying in-between the naive and the strict Efstathiou & Lemos (2018) definitions. As long as $t < s$, the sensitivity of the tension significance to the width of the TPD distribution is small.

For this toy model, we find very good agreement between the tension significance derived from our TPD approach and the standard PPD ansatz; see Fig. 4. In this case, the PPD can be obtained analytically as a convolution of Gaussian PDFs,

$$\ln \Pr(\hat{\boldsymbol{d}} \,|\, \boldsymbol{d}, \mathrm{H}_\alpha) = -\frac{1}{2}(\hat{\boldsymbol{d}} - \boldsymbol{d})^\tau \left[\mathbf{C} + \mathbf{C}^{\rm TPD}\right]^{-1} \times (\hat{\boldsymbol{d}} - \boldsymbol{d}) + \text{const.}, \tag{59}$$

while

$$\mathrm{I}_{\rm PPD} = 1 - \int_{V_N(\mu)} \mathrm{d}^N x \, \mathcal{N}\left(\boldsymbol{x}; \mathbf{0}, \mathbf{C} + \mathbf{C}^{\rm TPD}\right) = \frac{\Gamma\left(\frac{N}{2}, \frac{\mu^2}{2(s^2+t^2)}\right)}{\Gamma\left(\frac{N}{2}, 0\right)}. \tag{60}$$

As expected, the tension estimates agree as $t/s \to 0$. This can also be seen mathematically from equations (54) and (60) by taking the limits $\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{C}^{\rm TPD}\right) \to \delta_{\rm D}(\boldsymbol{x} - \boldsymbol{\mu})$ and $\mathbf{C} + \mathbf{C}^{\rm TPD} \to \mathbf{C}$. As t/s increases, the TPD estimate returns slightly less significant tension than the PPD version.

We refer the reader to Appendix C for a discussion on how our tension estimates are affected by measurement error, correlations between data, and sampling noise in the posterior.

4 DATA SET: KIDS-450

One of the primary targets for currently ongoing large-scale structure surveys such as KiDS, DES (DES Collaboration 2017), and the Hyper Suprime-Cam Survey (Mandelbaum et al. 2018) is to measure the weak gravitational lensing effect of the large-scale structure (see Kilbinger 2015 for a review and Bartelmann & Schneider 2001 for a more general introduction) in order to infer precise and accurate constraints on key cosmological parameters at low redshifts, $z \lesssim 1$, in contrast to the high-redshift constraints on those parameters from the CMB.

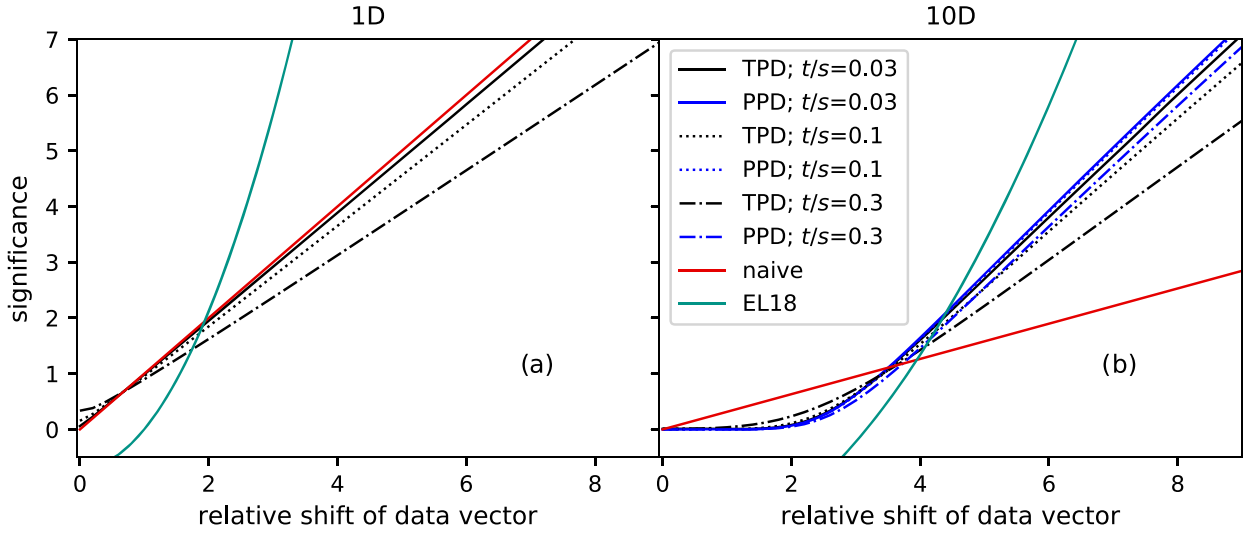

Figure 4. Tension significance criteria for a Gaussian toy model in (a) one dimension and (b) 10 dimensions. The significance $m\,\sigma$ is plotted as a function of the shift of the best-fitting model (i.e. the peak of the TPD or PPD) with respect to the data, μ, in units of the standard deviation of the data measurement error, s. Black (dark blue) lines correspond to the definition of tension based on the TPD (PPD) with different line styles showing the dependence on the TPD width, t (as given in the legend, in units of s). The light blue line follows the definition of Efstathiou & Lemos (2018). The red line is a naive criterion taken as the relative shift of the data vector, divided by $\sqrt{N}$, where N is the dimension of the distribution under consideration. In one dimension, the red line therefore marks a one-to-one relation (overlapping the blue solid line), which is closely approximated by the TPD and PPD definitions of significance as $t/s \to 0$.

For that purpose (several) thousand square degrees on the sky are observed in multicolour bands typically ranging from near-infrared to optical to measure galaxy positions and their shapes. The shape measurements are used to infer the gravitational shear, i.e. the tiny but coherent distortions imprinted on galaxy images due to the weak-lensing effect of the intervening large-scale structure through which light has to propagate before arriving at the observer. The measured shear and galaxy positions can then be used to build up the shear–shear two-point statistics, also termed cosmic shear. The real space two-point correlation functions (2PCF) or equivalently their power spectra are all related to the power spectrum of matter density fluctuations and therefore can be used to yield competitive constraints on the combination of the matter clustering amplitude, σ_8 – the root-mean-square dispersion of the density contrast measured in spheres of 8 h^{-1} Mpc on the sky – and the total matter density, Ω_m, i.e. $S_8 = \sigma_8\sqrt{\Omega_\mathrm{m}/0.3}$.

In the following application of the three tiers of consistency tests to data, we will use tomographic cosmic shear measurements from an intermediate data release based on 450 sq deg of imaging data from KiDS (Kuijken et al. 2015; Hildebrandt et al. 2017; Fenech Conti et al. 2017).[2]

The KiDS data are processed with THELI (Erben et al. 2013) and ASTRO-WISE (Begeman et al. 2013; de Jong et al. 2015). Shears are measured using *lens*fit (Miller et al. 2013), and photometric redshifts are obtained with BPZ (Benítez 2000) from point spread function matched photometry and calibrated using external overlapping spectroscopic surveys (see Hildebrandt et al. 2017 for details).

The KiDS-450 cosmic shear data were used in Hildebrandt et al. (2017) for a real space 2PCF analysis with the $\xi_+(\theta)$ and $\xi_-(\theta)$ estimators in four tomographic bins ($0.10 < z_1 \leq 0.30$, $0.30 < z_2 \leq 0.50$, $0.50 < z_3 \leq 0.70$, and $0.70 < z_4 \leq 0.90$) spanning angular scales $0.50 < \theta_+/\mathrm{arcmin} < 72$ and $4.2 < \theta_-/\mathrm{arcmin} < 300$. In addition to these fiducial scales, Hildebrandt et al. (2017) also defined a set of 'large' and 'small' angular scales for further systematic tests which we will also be using in the subsequent analysis. We summarize all angular scales and their abbreviations in Table 1 for convenience.

[2]The data are publicly available at http://kids.strw.leidenuniv.nl/sciencedata.php.

4.1 Data likelihood

The cosmological interpretation of the observed correlation-function estimators, $\xi_\pm^\alpha(\theta)$, is carried out in a Bayesian framework. For the estimation of cosmological model parameters, $\boldsymbol{p}$, we sample the posterior PDF by evaluating the likelihood

$$-2\ln\mathcal{L}(\boldsymbol{p}) = \sum_{\alpha,\,\beta} \boldsymbol{\Delta}_\alpha(\boldsymbol{p})\,(\mathbf{C}^{-1})_{\alpha\beta}\,\boldsymbol{\Delta}_\beta(\boldsymbol{p}), \tag{61}$$

where the indices α and β run over the *unique* tomographic redshift bin combinations. The analytical covariance matrix, $\mathbf{C}$, is calculated as outlined in Hildebrandt et al. (2017).

We note that Troxel et al. (2018) derived an update for that covariance with an improved shot-noise model, primarily incorporating previously neglected survey-boundary effects. This improves the goodness of fit of the fiducial model significantly with a χ^2 per degree of freedom close to unity (compare also to Table 2). However, for reasons of consistency we use subsequently the same model as employed in the original KiDS-450 analysis and the analysis of Efstathiou & Lemos (2018, see section 5.3), but we will comment on potential changes due to the updated covariance where applicable. To illustrate the correlations between angular scales but also between different redshift bin combinations, we show in Fig. 5 the correlation matrix of the covariance.

The components of the data vector are calculated as

$$\Delta_\alpha(\boldsymbol{p}) = \hat{\xi}_\pm^\alpha(\theta) - \xi_\pm^\alpha(\theta,\ \boldsymbol{p}), \tag{62}$$

where the hat denotes measurements extracted from the observations. The model predictions, $\xi_\pm^\alpha(\theta,\ \boldsymbol{p})$, for the $\xi_\pm$ correlation

Table 1. Sets of angular scales used in the analysis.

Abbreviation	Estimator	$\theta_{\rm min}$ (arcmin)	$\theta_{\rm max}$ (arcmin)	Number of θ-bins
Fiducial scales	ξ_+	0.50	72	7
Fiducial scales	ξ_-	4.2	300	6
Large scales	ξ_+	4.2	72	3
Large scales	ξ_-	4.2	300	6
Small scales	ξ_+	0.50	4.2	4
Small scales	ξ_-	–	–	0

Notes: The 'fiducial scales' and 'large scales' listed here correspond to the definitions in Hildebrandt et al. (2017) that were also used by Efstathiou & Lemos (2018, see section 5.3). Based on the 'large scales' definition, we construct the mutually exclusive 'small scales' set.

Table 2. Evidence ratios for various splits of the KiDS-450 $\xi_\pm$ data vector.

Data split	Model	B modes subtracted	χ^2	d.o.f.	$\ln(\mathcal{Z})$	$\log_{10}(\mathrm{R}_{01})$	Evidence for H_0 on Jeffreys' scale
–	H_0	No	160.44	123	-91.27 ± 0.09	–	–
Large versus small scales	H_1	No	154.73	116	-94.00 ± 0.11	1.19 ± 0.06	Strong
z-bin 3 versus all others	H_1	No	155.16	116	-95.48 ± 0.12	1.83 ± 0.06	Very strong
z-bin 4 versus all others	H_1	No	157.28	116	-96.93 ± 0.12	2.46 ± 0.06	Decisive
ξ_+ versus ξ_-	H_1	No	153.52	116	-94.71 ± 0.12	1.49 ± 0.06	Strong
–	H_0	Yes	137.00	123	-79.07 ± 0.08	–	–
Large versus small scales	H_1	Yes	139.45	116	-87.08 ± 0.11	3.48 ± 0.06	Decisive
z-bin 3 versus all others	H_1	Yes	129.83	116	-84.93 ± 0.12	2.55 ± 0.06	Decisive
z-bin 4 versus all others	H_1	Yes	115.95	116	-80.75 ± 0.12	0.73 ± 0.06	Substantial
ξ_+ versus ξ_-	H_1	Yes	140.40	116	-87.43 ± 0.11	3.63 ± 0.06	Decisive

Notes: The first column lists the split applied to the fiducial KiDS-450 data vector. The z-bin splits should always be read as, e.g. 'z-bin 3 (and all its CCs) versus all other z-bin correlations'. In the second column, we give the model that is used in the calculations. H_0 corresponds to the fiducial model using only one set of parameters, whereas H_1 uses separate parameter sets for each subsample of the split. The third column indicates whether or not the measured B modes were subtracted off the data vector. The remaining columns then list the χ^2 of the fit, the number of degrees of freedom (d.o.f.), the natural logarithm of the evidence $\mathcal{Z}$, the binary logarithm of the Bayes factor R_{01} and finally its qualitative interpretation on Jeffreys' scale. The latter must be read as evidence for the model H_0: 'there exists one common set of parameters that describe all data sets'.

functions as functions of angular separation, θ, between galaxies on the sky and between redshift-bin correlations, $z_\mu \times z_\nu$, are related to the tomographic E-mode convergence power spectrum, $C^{\rm EE}_{\mu\nu}(\ell)$, as a function of multipoles, ℓ, through Bessel functions of the first kind, $J_{0,4}$ (of order 0 for ξ_+ and of order 4 for ξ_-):

$$\xi^{\mu\nu}_\pm(\theta) = \frac{1}{2\pi}\int \mathrm{d}\ell\, \ell C^{\rm EE}_{\mu\nu}(\ell) J_{0,4}(\ell\theta). \tag{63}$$

The tomographic convergence power spectrum in the (extended) Limber approximation (Limber 1953, Kaiser 1992, LoVerde & Afshordi 2008) can be written as:

$$C^{\rm EE}_{\mu\nu}(\ell) = \int_0^{\chi_{\rm H}} \mathrm{d}\chi\, \frac{q_\mu(\chi) q_\nu(\chi)}{f^2_{\rm K}(\chi)} P_\delta\left(k = \frac{\ell + 0.5}{f_{\rm K}(\chi)}; \chi\right), \tag{64}$$

which depends on the comoving radial distance, χ, the comoving distance to the horizon, $\chi_{\rm H}$, the comoving angular diameter distance, $f_{\rm K}(\chi)$, and the 3D matter power spectrum, $P_\delta(k; \chi)$.

The weight functions, $q_\mu(\chi)$, depend on the lensing kernels and hence they are a measure of the lensing efficiency in each tomographic redshift bin, μ:

$$q_\mu(\chi) = \frac{3\Omega_{\rm m} H_0^2}{2c^2} \frac{f_{\rm K}(\chi)}{a(\chi)} \int_\chi^{\chi_{\rm H}} \mathrm{d}\chi'\, n_\mu(\chi') \frac{f_{\rm K}(\chi' - \chi)}{f_{\rm K}(\chi')}, \tag{65}$$

where $a(\chi)$ is the scale factor and the source redshift distribution is denoted as $n_\mu(\chi)\,\mathrm{d}\chi = n'_\mu(z)\,\mathrm{d}z$. It is normalized such that $\int \mathrm{d}\chi n_\mu(\chi) = 1$.

The observed shear correlation functions, $\xi^{\rm obs}_\pm$, are only a biased tracer of the cosmological signal encoded in the $\xi_\pm$ estimators due to intrinsic galaxy alignments:

$$\xi^{\rm obs}_\pm = \xi_\pm + \xi^{\rm II}_\pm + \xi^{\rm GI}_\pm. \tag{66}$$

Here, $\xi^{\rm II}_\pm$ measures the intrinsic ellipticity correlations between neighbouring galaxies (termed 'II') and $\xi^{\rm GI}_\pm$ encodes the correlations between the intrinsic ellipticities of foreground galaxies and the gravitational shear of background galaxies (termed 'GI'). We follow Hildebrandt et al. (2017) in modelling these effects and employ the non-linear modification of the tidal alignment model of intrinsic alignments (Hirata & Seljak 2004; Bridle & King 2007; Joachimi et al. 2011). The angular power spectra of the intrinsic alignments can be written as:

$$C^{\rm II}_{\mu\nu}(\ell) = \int_0^{\chi_{\rm H}} \mathrm{d}\chi\, \frac{n_\mu(\chi) n_\nu(\chi) F^2(\chi)}{f^2_{\rm K}(\chi)} P_\delta\left(k = \frac{\ell + 0.5}{f_{\rm K}(\chi)}; \chi\right), \tag{67}$$

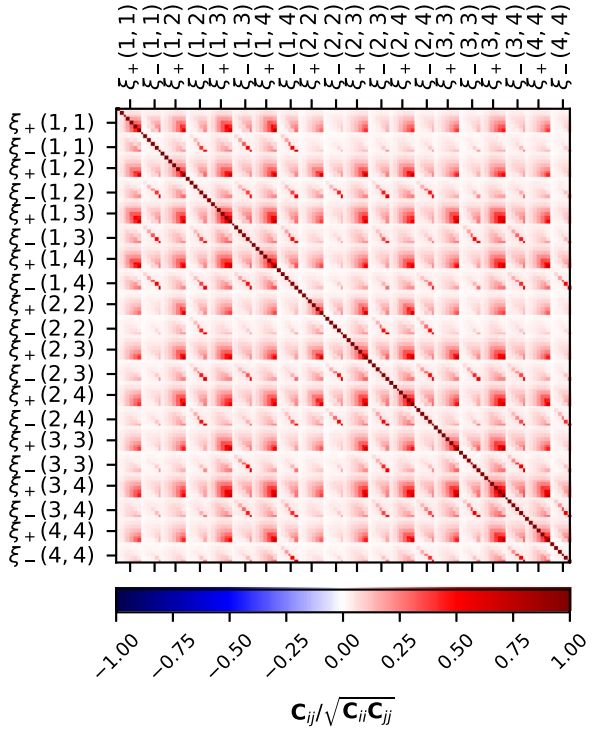

Figure 5. The correlation matrix of the $\xi_\pm$ correlation function covariance, $\mathbf{C}$, for all fiducial angular scales, θ, and tomographic bin combinations, $i \times j$ (see 'fiducial scales' in Table 1).

$$C^{\mathrm{GI}}_{\mu\nu}(\ell) = \int_0^{\chi_{\mathrm{H}}} \mathrm{d}\chi \, \frac{q_\nu(\chi) n_\mu(\chi) + q_\mu(\chi) n_\nu(\chi)}{f_{\mathrm{K}}^2(\chi)} F(\chi) P_\delta \left(k = \frac{\ell + 0.5}{f_{\mathrm{K}}(\chi)}; \chi \right), \tag{68}$$

with the lensing weight function, $q_\mu(\chi)$, from equation (65) and

$$F(\chi) = -A_{\mathrm{IA}} C_1 \rho_{\mathrm{crit}} \frac{\Omega_{\mathrm{m}}}{D_+(\chi)}. \tag{69}$$

The dimensionless amplitude A_{IA} allows us to rescale and vary the fixed normalization $C_1 = 5 \times 10^{-14}\, h^{-2} \mathrm{M}_\odot^{-1} \mathrm{Mpc}^3$ in the subsequent likelihood analysis. The critical density of the Universe today is denoted as ρ_{crit} and $D_+(\chi)$ is the linear growth factor normalized to unity today.

Another astrophysical effect that needs to be taken into account is baryon feedback, i.e. modifications of the matter distribution at small scales, for example, due to active galactic nucleus (AGN) feedback (e.g. Semboloni et al. 2011; Semboloni, Hoekstra & Schaye 2013). The full physical description of baryon feedback is not established yet and different 'recipes' exist usually based on hydrodynamical simulations. The effect of baryon feedback is typically quantified as a bias with respect to the dark-matter-only matter power spectrum, P_δ (e.g. Semboloni et al. 2013; Harnois-Déraps et al. 2015):

$$b^2(k, z) \equiv \frac{P_\delta^{\mathrm{mod}}(k, z)}{P_\delta^{\mathrm{ref}}(k, z)}, \tag{70}$$

where P_δ^{mod} and P_δ^{ref} denote the power spectra with and without baryon feedback, respectively.

In Hildebrandt et al. (2017), the baryon feedback model included in HMcode by Mead et al. (2015, 2016) was used. However, this module for the non-linear matter power spectrum is not yet available for the Boltzmann-code CLASS[3] (Blas, Lesgourgues & Tram 2011; Audren & Lesgourgues 2011). Therefore, we use here the HALOFIT algorithm within CLASS (including the Takahashi et al. 2012 recalibration) and add the baryon feedback model through the fitting formula for baryon feedback from Harnois-Déraps et al. (2015) based on the AGN model from the OverWhelmingly Large Simulations (Schaye et al. 2010, van Daalen et al. 2011):

$$b^2(k, z) = 1 - A_{\mathrm{bary}}[A_z \mathrm{e}^{(B_z x - C_z)^3} - D_z x \mathrm{e}^{E_z x}], \tag{71}$$

where $x = \log_{10}(k/h\,\mathrm{Mpc}^{-1})$ and the terms A_z, B_z, C_z, D_z, and E_z are feedback model-dependent functions of the scale factor $a = 1/(1 + z)$. We refer the reader to Harnois-Déraps et al. (2015) for the specific functional forms and constants. Moreover, we introduce a free amplitude, A_{bary}, to marginalize over while fitting for the cosmological parameters.

In the likelihood analysis, we assume a cosmological model with spatially flat geometry and use the same set of key cosmological parameters and priors as in Hildebrandt et al. (2017): $\Omega_{\mathrm{cdm}}h^2$, $\ln(10^{10}A_{\mathrm{s}})$, $\Omega_{\mathrm{b}}h^2$, n_{s}, h, i.e. the amplitude of the primordial power spectrum A_{s}, the value h of the Hubble parameter today divided by $100\,\mathrm{km\,s^{-1}Mpc^{-1}}$, the cold dark matter density $\Omega_{\mathrm{cdm}}h^2$, the baryonic matter density $\Omega_{\mathrm{b}}h^2$, and the exponent of the primordial power spectrum n_{s}. In addition to these key cosmological parameters, we add the free amplitude parameters A_{IA} and A_{bary} for the intrinsic alignment and baryon feedback model, the former again in the same prior range as in Hildebrandt et al. (2017). We emphasize that the likelihood pipeline used here is independent of the cosmology pipeline used in Hildebrandt et al. (2017) with the additional difference in the baryon feedback model and the prior on its amplitude, A_{bary}. However, we find that the impact of that is negligible and our pipeline recovers a $\chi^2_{\mathrm{min}} = 160.4$ and $S_8 = 0.756 \pm 0.037$ in the fiducial joint set-up in comparison to $\chi^2_{\mathrm{min}} = 162.5$ and $S_8 = 0.745 \pm 0.039$ as found in Hildebrandt et al. (2017).

For an efficient evaluation of the likelihood $\mathcal{L}$, we employ the nested sampling algorithm MULTINEST (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009; Feroz et al. 2013).[4] Conveniently, its PYTHON-wrapper PYMULTINEST (Buchner et al. 2014) is included in the framework of the cosmological likelihood sampling package MONTE PYTHON (Audren et al. 2013) with which we derive all cosmology-related results in this analysis.[5]

We will refer to the posterior samples derived with the nested sampling algorithm as an MCMC. Moreover, we note that the weights connected to each MCMC sample are always propagated consistently in the subsequent analysis. For example, when we refer to the mean of a quantity, we calculate its *weighted* mean.

5 APPLICATION OF CONSISTENCY TESTS TO KIDS-450

In the following, we assess the internal consistency of the fiducial KiDS-450 correlation function analysis making use of the tests established in Section 2 and tested in more detail in Appendix A. The KiDS-450 cosmic shear data present an excellent test case for assessing consistency in a highly correlated data set and is also motivated by the following findings: Hildebrandt et al. (2017)

[3] Version 2.5.0 from https://github.com/lesgourg/class_public
[4] Version 3.8 from http://ccpforge.cse.rl.ac.uk/gf/project/multinest/
[5] Version 2.2.1 from https://github.com/baudren/montepython_public

reported in their section 6.5 a shift to lower S_8 values with respect to the fiducial results when including only large angular scales in the ξ_+ measurements (see Table 1), as well as when applying large-scale cuts to both ξ_+ and ξ_- (Joudaki et al. 2017). Since this shift to lower S_8 values is also observed in the quadratic estimator analysis of Köhlinger et al. (2017), which in general uses larger scales than the correlation function analysis (see fig. C1 in Köhlinger et al. 2017), this may hint at inconsistencies between large and small angular scales in the data. Therefore, the first split of the fiducial data vector consists of two mutually exclusive subsets containing either the large or small angular scales, respectively (see Table 1).

Efstathiou & Lemos (2018) found that the scaling of some tomographic bin combinations might be inconsistent, reporting the largest inconsistencies for z-bins 3 and 4 ($0.50 \leq z_3 < 0.70$ and $0.70 \leq z_4 < 0.90$; see also Section 5.3). Joudaki et al. (2017) also found hints for an inconsistency in the source redshift distribution of z-bin 3 (see their appendix A) and van Uitert et al. (2018) show in a combined analysis of cosmic shear, galaxy–galaxy lensing and angular galaxy clustering that the data preferred to shift z-bin 3 by $\mathrm{d}z = -0.061^{+0.010}_{-0.039}$, while for the other z-bins no significant shifts are observed.

In addition to that, a comparison of the source redshift distributions derived with the direct calibration method ('DIR') and a cross-correlation method ('CC'; see Hildebrandt et al. 2017 for details) reveals the largest deviations between these two methods for z-bin 3. Therefore, we investigate the consistency of the redshift scaling with a split of the fiducial data into mutually exclusive subsets containing only z-bin 3 (and all its CCs) versus all other tomographic bin combinations. This check is repeated again for z-bin 4. We intentionally do not use the lower redshift bins 1 and 2 for this test due to the lower signal-to-noise ratio (S/N) in these bins compared to z-bins 3 and 4.

Finally, Hildebrandt et al. (2017) present in their appendix D6 a decomposition of the fiducial correlation function data into E and B modes. A non-zero detection of B modes indicates that residual systematics are present in the data. If the systematics produce E and B modes with equal strength, it can be mitigated according to equation (A2). Although mitigating this effect was shown to not affect the cosmological results significantly, we split the fiducial data vector into mutually exclusive ξ_+ and ξ_- subsets to assess the significance of the measured small-scale B modes in the KiDS-450 data. Moreover, we also repeat all consistency checks for the data splits mentioned above for a data vector from which we subtract (two times) the measured B modes from the ξ_+ correlation functions (implicitly assuming that the systematic generates equal power in E and B modes).

5.1 Consistency in posterior parameter space

Following Section 2.1, we perform the analysis as follows: we use the KiDS-450 $\xi_\pm$ data vector and the KiDS-450 covariance matrix within the fiducial scales (see Table 1) as the input for the joint MCMC run (i.e. the numerator of equation 3) corresponding to the model H_0: 'there exists one common set of parameters that describe all data sets' and sample the likelihood in the same parameters and prior ranges as presented in Hildebrandt et al. (2017) with the caveats discussed in Section 4.1.

For the split MCMC run (i.e. the denominator of equation 3) which tests now the model H_1: 'there exist two separate parameter sets that each describe one subset of the data',[6] we split the fiducial KiDS-450 data vector according to the systematic we want to test. For example, to detect a shift in the source redshift distribution of z-bin 3, we split the data vector $\boldsymbol{d}_{\rm tot}^{\rm data}$ into one set $\boldsymbol{d}_{\rm a}^{\rm data}$ containing only z-bin 3 (and all its CCs) and the mutually exclusive set $\boldsymbol{d}_{\rm b}^{\rm data}$ containing all other unperturbed z-bins (and their CCs), thus $\boldsymbol{d}_{\rm tot}^{\rm data^\intercal} = \{\boldsymbol{d}_{\rm a}^{\rm data^\intercal}, \boldsymbol{d}_{\rm b}^{\rm data^\intercal}\}$.

It is important to note that both subsets, $\boldsymbol{d}_{\rm a}^{\rm data}$ and $\boldsymbol{d}_{\rm b}^{\rm data}$, of the split data set are still coupled through the full covariance (which is the same as used in the joint MCMC run with $\boldsymbol{d}_{\rm tot}^{\rm data}$ by construction), but as mentioned in Section 2.1 we keep all cosmology-dependent calculations as well as all cosmological and nuisance parameters separated in the likelihood analysis. In total, the joint MCMC uses the five cosmological and two nuisance parameters as listed in Section 4.1 and hence the split MCMC uses 14 parameters for typically $(n_{\theta+} + n_{\theta-})\, n_{\rm z}(n_{\rm z}+1)/2$ data points (e.g. for the 'fiducial scales' from Table 1 that corresponds to $n_{\rm z} = 4$, $n_{\theta+} = 7$, and $n_{\theta-} = 6$, i.e. 130 data points in total).

While sampling the joint and split MCMCs for all four splits of the data vector as listed in Table 2, we also calculate the evidences and the respective Bayes factors. These reveal no significant tension for any of the data splits and instead yield at least 'strong' (large versus small scales and ξ_+ versus ξ_-) to 'decisive' (z-bin 4 versus all others) evidence on Jeffreys' scale for the fiducial model H_0: 'there exists one common set of parameters that describes all data sets'. Subtracting off the measured small-scale B modes from the data vector generally strengthens the evidence for the fiducial model with the exception of splitting the data into the subsets containing z-bin 4 and all its CCs versus all other tomographic bin combinations ('z-bin 4 versus all others'). For this split, the evidence decreases from 'decisive' to 'substantial' which we interpret as a sign that an inconsistency in z-bin 4 becomes more pronounced once the B modes are subtracted off. We note though that based on the sensitivity analysis performed in Appendix A1, we find the Bayes factor test only to be a necessary criterion for consistency, not a sufficient one (see also Raveri & Hu 2018). This is due to the prior volume which has a significant impact on the Bayes factor, especially when most parameters are prior-driven. Wide prior ranges on parameters that are only weakly constrained by the data will then lower the evidence in general (compare also to Fig. 3a). Moreover, it quantifies the general goodness of fit of a model rather than tension.

Hence, we proceed with the second tier of consistency tests, for which we compare the differences between the duplicate parameter sets of the split MCMC run. Although all seven primary parameters are duplicated in that run, we focus here only on the duplicates of two derived cosmological parameters and one primary nuisance parameter. In particular, those are the parameter constrained best by cosmic shear, i.e. S_8, and the total matter density, $\Omega_{\rm m}$, as these two parameters set the amplitude and the tilt of the cosmic shear signal. The third parameter is the amplitude of the intrinsic alignment model, $A_{\rm IA}$. This nuisance parameter is of particular interest because it is degenerate with the other two derived cosmological parameters and hence it also affects the amplitude and tilt of the cosmic shear signal.

Indeed, we observe for the 2D projections of these key parameter differences shown in Fig. 6 similar trends as seen in the Bayes factor analysis: for example for the z-bin 3/4 splits (Figs 6b and d) the 68 per cent and 95 per cent (inner and outer) credibility contours for which the B modes are subtracted off (dotted contours) are more

[6] Note that this is the more specific version of H_1 as given in Section 2.1.

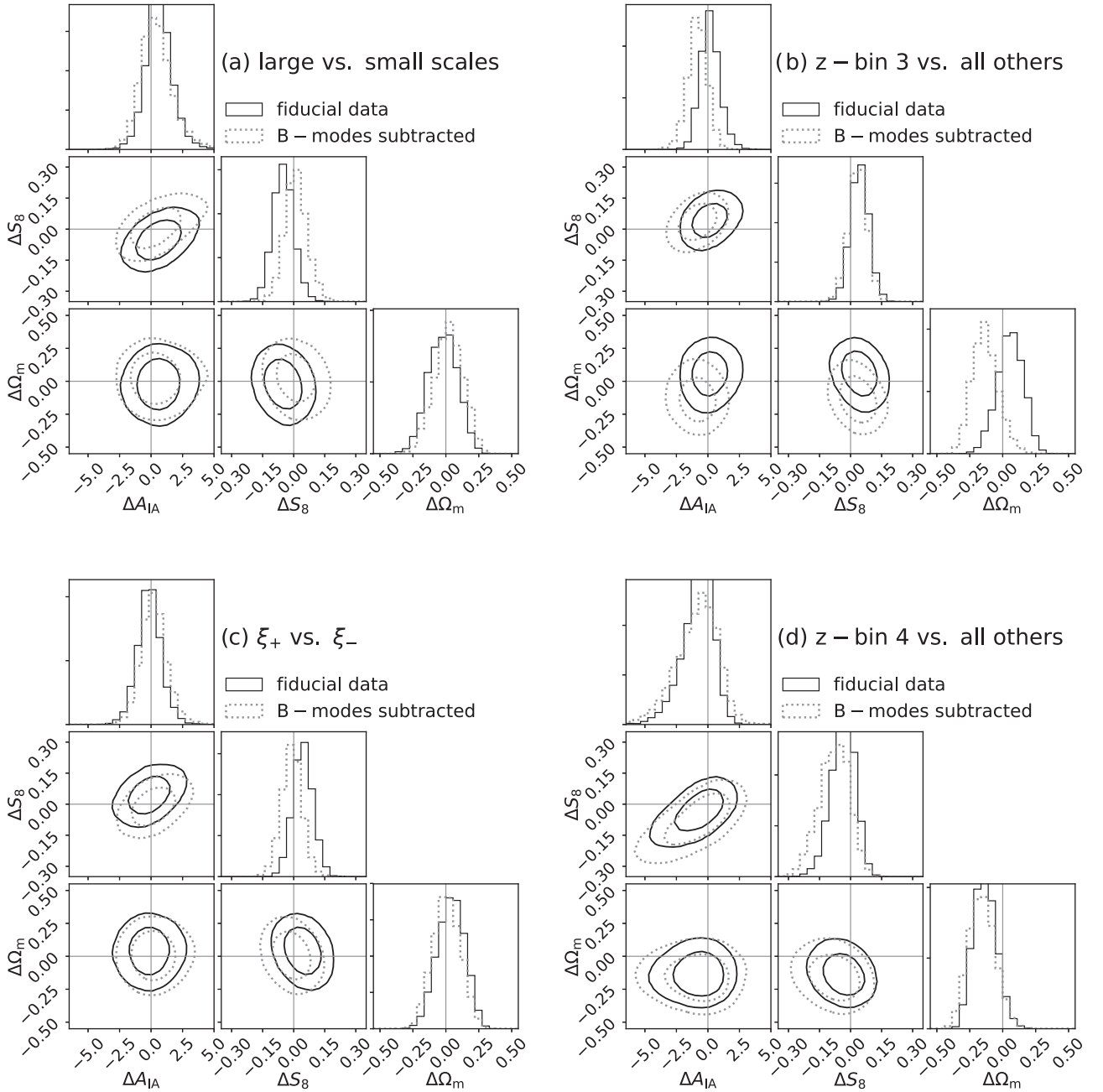

Figure 6. Duplicate parameter differences from the split MCMC run for 2D projections of key parameters for the following splits of the KiDS-450 data set: (a) large versus small scales (angular) scales, (b) z-bin 3 (and all its CC) versus all other z-bin combinations, (c) ξ_+ versus ξ_-, and (d) z-bin 4 (and all its CC) versus all other z-bin combinations. The covariance between the exclusive sets of the split MCMC run is fully taken into account in the parameter inference. The solid black contours (the inner/outer one corresponding to the 68/95 per cent credibility interval) show the biases (i.e. offsets with respect to the cross hairs) in the parameter projections derived using the fiducial KiDS-450 data vector. In contrast to that the dashed blue contours show the impact of removing the measured B modes from the data vector on the splits into subsets.

biased than the contours for the fiducial KiDS-450 data vector (solid contours).

For the other data splits though we observe that the $\sim 1\,\sigma$-level biases decrease once the B modes are subtracted off. In general, all conclusions drawn from the Bayes factor results are strongly supported by the key parameter differences: there are no signs for strong residual systematics and biases in the posterior parameters are ranging at most between $\sim 1\,\sigma$ to $\leq 2.70\,\sigma$ for all parameter projections, the strongest biases occurring for the B-mode subtracted z-bin 4 split.

Following the method outlined in Section 2.2, we also quantify the significances for all 2D parameter projections in Table 3. Moreover, we also calculate the significances for tension over the full three key-parameter subspace. These also support the conclusions from the Bayes factor: subtracting off the B modes from the data vector increases the tension in case of the z-bin 4 split, i.e.

Table 3. Significances for duplicate parameter differences.

Data split	B modes subtracted	$\Delta\{S_8, \Omega_m, A_{IA}\}$	$\Delta\{S_8, \Omega_m\}$	$\Delta\{S_8, A_{IA}\}$	$\Delta\{\Omega_m, A_{IA}\}$
Large versus small scales	No	$1.32\,\sigma$	$0.72\,\sigma$	$1.25\,\sigma$	$0.21\,\sigma$
z-bin 3 versus all others	No	$0.80\,\sigma$	$1.00\,\sigma$	$0.51\,\sigma$	$0.25\,\sigma$
z-bin 4 versus all others	No	$0.92\,\sigma$	$1.33\,\sigma$	$0.11\,\sigma$	$1.08\,\sigma$
ξ_+ versus ξ_-	No	$0.77\,\sigma$	$0.77\,\sigma$	$0.69\,\sigma$	$0.08\,\sigma$
Large versus small scales	Yes	$0.00\,\sigma$	$0.05\,\sigma$	$0.01\,\sigma$	$0.05\,\sigma$
z-bin 3 versus all others	Yes	$1.46\,\sigma$	$0.95\,\sigma$	$1.55\,\sigma$	$1.37\,\sigma$
z-bin 4 versus all others	Yes	$2.42\,\sigma$	$2.71\,\sigma$	$0.35\,\sigma$	$2.04\,\sigma$
ξ_+ versus ξ_-	Yes	$0.06\,\sigma$	$0.02\,\sigma$	$0.15\,\sigma$	$0.04\,\sigma$

Notes: The significances for tension in the listed 2D projections of the key parameter differences and the full three-parameter subspace (third column) are derived as described in Section 2.2. Moreover, the results for the 2D projections can directly be compared to the contours shown in Fig. 6.

from $0.92\,\sigma$ to $2.42\,\sigma$, while it decreases the tension significantly for the 'large versus small scales' and 'ξ_+ versus ξ_-' splits, from $\leq 1.32\,\sigma$ to $\sim 0\,\sigma$. In contrast to that though, the overall tension in the z-bin 3 split decreases according to the Bayes factor, but increases from $0.80\,\sigma$ to $1.46\,\sigma$ when subtracting off the B modes. However, that is due to the dimensionality of the parameter spaces involved in each tension estimator: for the Bayes factor the full parameter space is used, whereas the significances in Table 3 are only calculated for the subspaces of key parameters. In summary, we do not find hints for significant tension (i.e. $\geq 3\,\sigma$) for any of the tests taking place in parameter space.

5.2 Consistency in the data domain

Having investigated potential residual systematics for the four data splits in posterior parameter space, we now turn to the data domain and directly look at the ξ_+ correlation functions per unique tomographic bin combination $z_i \times z_j$ in the four panels of Fig. 7; the ξ_- correlation functions can be found in Appendix D (Fig. D1). The black points with error bars are the KiDS-450 data (the error bars are derived from the diagonal elements of the fiducial covariance matrix) and the red and blue/cyan points with error bars represent the means with their 68 per cent credibility intervals derived from the joint and split TPDs, respectively. In general, the joint TPDs (red) can be interpreted as a best-fitting model over all panels (also including the ξ_- correlation functions), whereas the blue and cyan points are based on the two separate sets of cosmological and nuisance parameters and usually yield slightly closer matches to the data (e.g. for the small versus large angular scales in Fig. 7a). We caution the reader against performing a 'χ–by–eye' estimate on the significance of any apparent feature since the correlations between angular scales and tomographic bin combinations are non-trivial (see also Fig. 5).

Hence, we proceed to compare the joint and split TPDs quantitatively to the (multivariate Gaussian) data distribution as outlined in Section 2.3 in order to assign significances to the trends visible in Fig. 7. The results for all four data splits (from left to right) are shown in Fig. 8(a). It is interesting to point out that, when we calculate the significances for the full data vector including all tomographic bin combinations and the fiducial angular scales corresponding to all panels in Figs 7(a)–(d) (and including the corresponding ξ_- correlation function panels in Figs D1a–d), we observe an almost constant significance level of ~ 2.0 to $\lesssim 2.5\,\sigma$ for any of the four data splits (grey crosses). This generally indicates that the theory model is only a moderately good fit to the data, which can also be read off from the χ^2-values given in Table 2.

Looking then at the significances for each subset of the splits (i.e. estimating the significances only for the panels containing either the light or dark blue points in Figs 7 and D1), we find that the subsets containing 'large (angular) scales', 'z-bin 4 (and all its CCs)' or 'ξ_-' also produce significances just below $\sim 2.5\,\sigma$. However, these significances are not dependent on whether the joint or split TPDs (circles and crosses) were used, hinting at a general mismatch between theory and data which is not dependent on the particular data split. Subtracting off the B modes from the fiducial data vector (indicated with 'no B' in Fig. 8a), however, lowers the significances for the 'fiducial' case, as expected from the improved χ^2-values (see Table 2). Hence, this consistency test quantifies the overall goodness of fit of the model (see also Appendix A3).

We note that the mismatch between theory and data throughout all splits flagged by this goodness-of-fit estimator can be explained by the update of the covariance matrix by Troxel et al. (2018) which was not applied in these tests in order to be consistent with the original KiDS-450 analysis and the one carried out by Efstathiou & Lemos (2018, see section 5.3). As mentioned already in Section 4.1, these authors propose to use an improved shot-noise model, mainly incorporating previously neglected survey-boundary effects, when calculating the covariance matrix. They further show that this update improves the goodness of fit of the fiducial model significantly and reduces the χ^2 per degree of freedom of currently ~ 1.30 (Table 2) to a value close to unity.

In order to get an estimate of the tension due to residual systematics in the data, we employ the second TPD-based estimate comparing the joint TPD directly to the split TPD assigning significances by fitting their difference to zero (see Appendix B for details on how the errors are estimated also accounting for all correlations). The results for all four data splits are shown (from left to right) in Fig. 8(b). First, we notice that all these significances are lower than the ones presented in the corresponding Fig. 8(a) for the goodness-of-fit estimator as expected. Moreover, we do not observe any clear trends between the different subsets (such as 'large versus small angular scales'). It is interesting to point out that, when subtracting off the measured B modes from the data vector, all significances decrease further except for the split 'z-bin 4 versus all others'. Although this is not the case for the first TPD-based estimator (cf. Fig. 8a), we would have expected such a behaviour based on our previous analysis in the posterior parameter space (cf. Fig. 6d).

In Appendix D, we show the significances for both TPD-based estimators for all ξ_+ and ξ_- correlation functions per unique

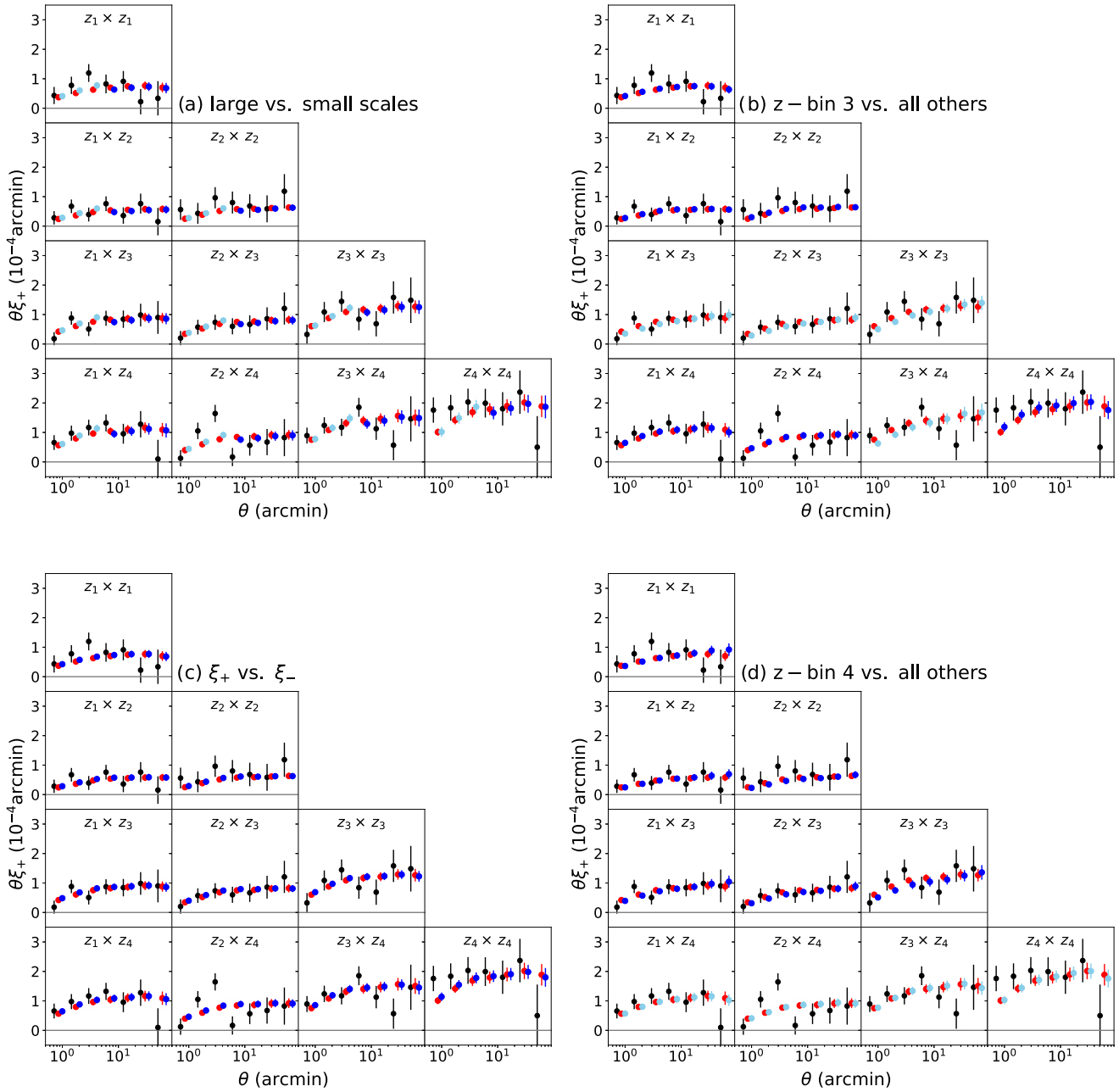

Figure 7. KiDS-450 data vector (black points) and TPD means (red and blue points) for the ξ_+ estimator as a function of angular scale per redshift bin combination. The TPDs are based on the joint (red points; the same in all panels) and split (dark and light blue points) cosmological and nuisance parameters from MCMC runs fit to the data vector. The panels consider various mutually exclusive splits of the data vector (a) large versus small scales, (b) z-bin 3 (and all its CCs) versus all other redshift correlations, (c) ξ_+ versus ξ_-, and (d) z-bin 4 (and all its CCs) versus all other redshift correlations. Error bars on the means are derived from the 68 per cent credibility interval around the mean. The erro rbars for the data are based on the diagonal of the covariance matrix. The corresponding plots for the ξ_- estimator can be found in Appendix D (Fig. D1).

tomographic bin combination (Figs D2 and D3). It is interesting to point out that for the TPD to data comparison in Fig. D2 the highest significance for tension is found in ξ_+ for the tomographic bin combination $z_2 \times z_4$ (at $\sim 2.6\,\sigma$) and in ξ_- for $z_1 \times z_2$ (at $\sim 1.9\,\sigma$) *independent* of the splits applied and also independent of whether the joint or split TPD were used. We do not observe a similar behaviour for the second TPD-based estimator which suggests that the data in the $z_2 \times z_4$ and $z_1 \times z_2$ tomographic bin combinations are the major causes driving the significances for the total data set or the two subsets as depicted in Fig. 8. Figs 7 and D1 also show that in the $z_2 \times z_4$ and $z_1 \times z_2$ panels the data points show large/the largest deviations with respect to the joint or split TPDs. Subtracting off the measured B modes again decreases all significances except for the split into the 'z-bin 4 versus all (other z-bin combinations)' subsets, as noted already above.

In summary, we remark that the first TPD-based estimator, comparing the joint and split TPDs to the data distribution, flags a general inconsistency between the model and the data at $\sim 2.5\sigma$

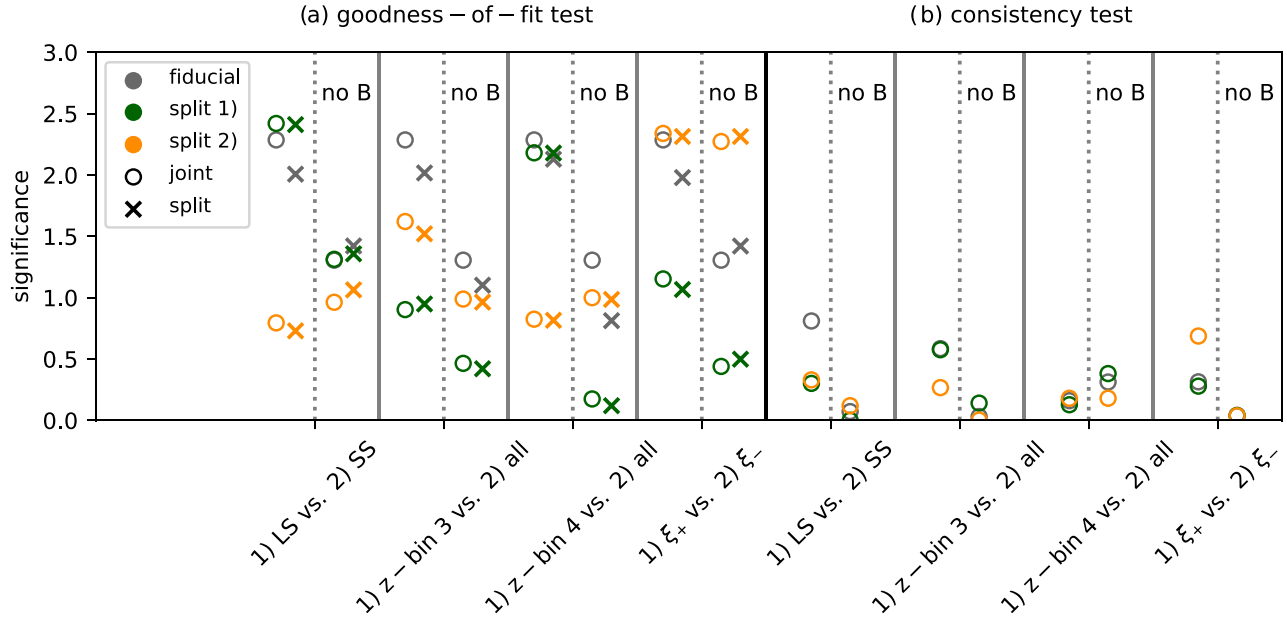

Figure 8. Significance as derived with the two TPD-based consistency estimators for different splits of the KiDS-450 data vector. (a) The significances for the goodness of fit are estimated by comparing the data distribution to the joint (circles) and split (crosses) TPDs. Grey symbols indicate that all fiducial angular scales (Table 1) are used in the significance estimation, whereas green and orange symbols correspond to one of the mutually exclusive subsets as indicated in the labels and along the x-axis. Shown are also the results when the measured B modes are subtracted from the data vector ('no B'). (b) The significances for tension are estimated by comparing the differences between the joint and the split TPDs.

for the 'fiducial' scales (see also Troxel et al. 2018) and at $\sim 2.0\,\sigma$ depending on which split is applied. Decomposing the data vector further into the tomographic correlation functions reveals that the major drivers for the bad goodness of fit arise from the $z_2 \times z_4$ (for ξ_+) and $z_1 \times z_2$ (for ξ_-) tomographic bin combinations. In comparison, the second TPD-based estimator, comparing the differences between the joint and the split TPDs directly, yields lower significances for tension and is qualitatively consistent with the results of the analysis in posterior parameter space in Section 5.1.

5.3 Comparison with Efstathiou & Lemos (2018)

Here, we provide a link from our three tiers of consistency checks to the one presented by Efstathiou & Lemos (2018). These authors use a cross-validation approach for which they split the fiducial data vector into mutually exclusive subsets $\boldsymbol{x}^D$ and $\boldsymbol{y}^D$ (for the cases presented here their choice and our choice of subsets coincides). For the larger of both subsets, $\boldsymbol{y}^D$, they infer best-fitting cosmological and nuisance parameters through an MCMC evaluation. For the best-fitting parameters, the corresponding full theory vector, $\{\boldsymbol{x}_{\rm model}, \boldsymbol{y}_{\rm model}\}$, is calculated and used to make a prediction for the vector $\boldsymbol{x}^D$ conditional on the fit to $\boldsymbol{y}^D$:

$$\boldsymbol{x}^{\rm cond} = \boldsymbol{x}_{\rm model} + \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\,(\boldsymbol{y}^D - \boldsymbol{y}_{\rm model}), \tag{72}$$

where the subscripts to the covariance $\mathbf{C}$ denote the submatrices corresponding to the respective selection from the data vector. The covariance of $\boldsymbol{x}^{\rm cond}$ is given as

$$\mathbf{C}_{xx}^{\rm cond} = \mathbf{C}_{xx} - \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}, \tag{73}$$

which can be used to calculate a conditional χ^2,

$$\chi^2_{\rm cond} = (\boldsymbol{x}^D - \boldsymbol{x}_{\rm model})^{\rm T}(\mathbf{C}_{xx}^{\rm cond})^{-1}(\boldsymbol{x}^D - \boldsymbol{x}_{\rm model}). \tag{74}$$

The significance of tension is then defined as the number of standard deviations by which $\chi^2_{\rm cond}$ deviates from the length N_x of the vector $\mathbf{x}^D$:

$$N_{\sigma_{\rm cond}} = (\chi^2_{\rm cond} - N_x)/\sqrt{2N_x}. \tag{75}$$

We emphasize that this definition of significance is generally more conservative than the one used in our approach (cf. Fig. 4b and Section 3 for details). Moreover, the definition of $N_{\sigma_{\rm cond}}$ approximates a χ^2-distribution with a Gaussian, which fails especially for smaller degrees of freedom and for the tails of the χ^2-distribution.

As discussed in Section 4.1, our likelihood pipeline is independent of the one used in Hildebrandt et al. (2017) and Efstathiou & Lemos (2018). Therefore, we repeat their calculations here with the caveat that we do not include the propagation of the model uncertainty in these repeated calculations that was incorporated into later versions of Efstathiou & Lemos (2018) and found to have only a small effect. The original numbers and our repeated results are listed in the first two columns of Table 4. With the exception of 'minus ξ_-', we reproduce the results of Efstathiou & Lemos (2018) well (our results are expected to yield slightly higher significances due to not propagating the model uncertainty). For the remainder of the comparison of the two approaches, we will refer to our repeated calculations when referring to the cross-validation approach unless stated otherwise.

In Figs 9(a) and (b), we give a visual impression of the cross-validation approach for the 'minus z-bin 3/4' cases. The KiDS-450 data vector (red points with error bars) is shown for all redshift bin combinations containing z-bin 3 and those containing z-bin 4 compared to the expected model vector conditional on the rest of the data, $\boldsymbol{x}^{\rm cond}$. The grey bands mark the $\pm 1\,\sigma$ and $\pm 2\,\sigma$ intervals around the expected model vector, $\boldsymbol{x}^{\rm cond}$, derived from the diagonal components of the conditional covariance matrix (equation 73). Fig. 9(a) shows that redshift bin combinations including z-bin 3 prefer a lower amplitude than the rest of the data. This problem is particularly apparent for ξ_- (lower panel) for the $z_3 \times z_3$ and

Table 4. Comparison of significances between the approach presented here and in Efstathiou & Lemos (2018).

Data used	E&L (as published)	E&L (repeated)	Data versus best fit (i.e. δ-TPD) (cross-validation)	Data versus TPD (cross-validionat)	Data versus TPD (joint and split)	TPD versus TPD (joint − split)
Minus z-bin 3	2.60	2.78	1.50	1.15	0.90 (joint), 0.95 (split)	0.58
Minus z-bin 4	3.52	3.58	2.48	2.28	2.18 (joint), 2.18 (split)	0.13
Minus ξ_-	2.71	1.95	2.45	2.36	2.34 (joint), 2.31 (split)	0.69
Minus ξ_+	1.20	1.66	2.23	1.86	1.15 (joint), 1.07 (split)	0.28

Notes:. The first column indicates which data was used in the MCMC evaluation (i.e. the full data vector 'minus ...'). The next two columns quote the numbers for the cross-validation approach as published in Efstathiou & Lemos (2018) and as recalculated with the likelihood pipeline used here (see Section 4.1 for details). The next two columns list the results from an approach linking the cross-validation significance to the symmetric TPD-based one used here (see the text for details). The last two columns report the significances from the TPD-based estimators as shown in Fig. 8. We quote here the numbers using only the parts of the data containing, for example, z-bins 3 and 4 in the calculation of the significances (i.e. the green symbols in that figure).

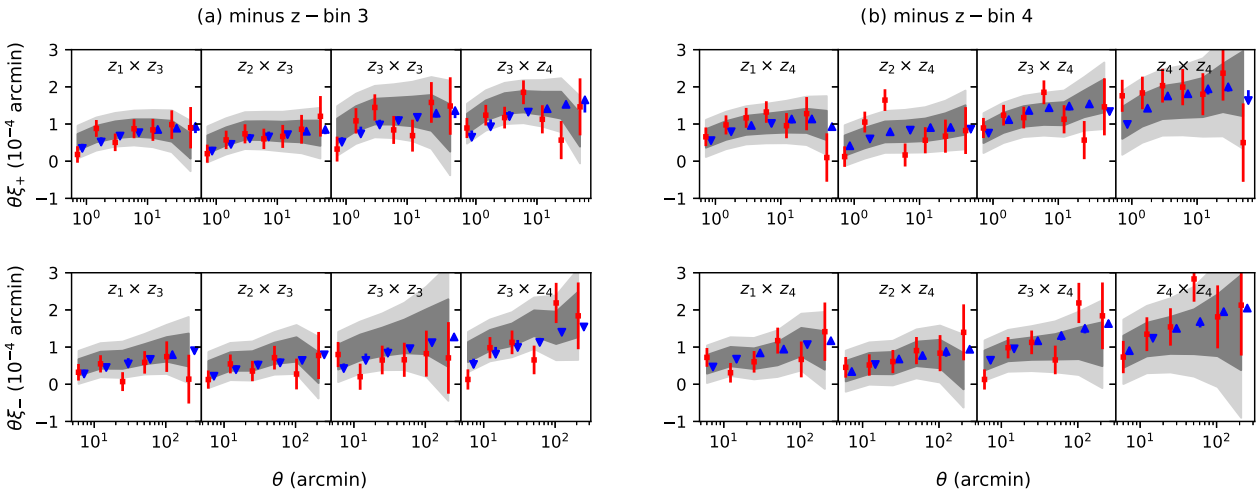

Figure 9. The ξ_+ (upper panel) and ξ_- (lower panel) correlation functions for all tomographic bin combinations $z_i \times z_j$ containing (a) z-bin 3 and (b) z-bin 4 (red points with error bars). The grey bands show the $\pm 1\,\sigma$ (dark grey) and $\pm 2\,\sigma$ (light grey) ranges allowed by the fits to the rest of the data (not containing z-bin 3 or 4). The blue arrows (displaced along the θ-axis for better visibility) are drawn from the mean value of the joint TPD to the split TPD and hence provide a qualitative measure for the tension at that particular θ-scale according to our TPD-based tension estimator.

$z_3 \times z_4$ combinations. These two redshift bin combinations carry a high weight in fits to the full data vector, yet they appear to be inconsistent at $\sim 2.8\,\sigma$ with the rest of the data according to this estimator (equation 75).

The situation appears to be more severe for the redshift bin combinations containing z-bin 4 since those produce a mismatch between expected model vector and the rest of the data at $\sim 3.6\,\sigma$. Both figures agree qualitatively with the conclusions presented by Efstathiou & Lemos (2018) and to link those visually to our TPD approach, we further include in each panel (blue) arrows pointing from the mean of the joint TPD to the mean of the split TPD. Hence, the length of the arrows is a qualitative measure for the strength of the tension at that particular θ-scale according to the TPD-based tension estimator (the longer the arrow in $\pm y$-direction the stronger the tension).

To also provide a link from our definition of significance to the Efstathiou & Lemos (2018) cross-validation approach, we interpret the best-fitting model vector $\boldsymbol{x}_{\rm model}$, obtained from fitting only the $\mathbf{y}^D$-part in the MCMC as a TPD with zero width, i.e. a Dirac δ-distribution. Then, we estimate a significance by comparing the 'δ-TPD' to the data distribution as outlined in Section 2.3 instead of using the equations of the cross-validation approach. The results for this are listed in the third column of Table 4. Although the significances for the cases 'minus z-bin 3' and 'minus z-bin 4' decrease with respect to the repeated Efstathiou & Lemos (2018) results (second column), they increase for the cases 'minus ξ_-' and 'minus ξ_+', thus reflecting the impact of the choice of the significance criteria. In the fourth column of Table 4, we also account for the model uncertainty by employing all model vectors sampled in the MCMC, i.e. the full TPD with a finite width. As expected, this decreases the significances further by $\sim 0.1\,\sigma$ to $\sim 0.4\,\sigma$ (from the third to fourth column).

We now drop the cross-validation approach entirely and switch to our previous symmetric approach and use the joint and split TPDs to estimate significances for the last two columns of Table 4. In particular, the fifth column lists the significances for comparing both joint and split TPDs individually to the data distribution using only the $\boldsymbol{x}^D$ part of the data in the significance estimates (compare also to the green/orange circles/crosses in Fig. 8). With respect to the previous four columns, all significances decrease

even further. This is expected because the joint and split MCMC runs are more constrained due to being coupled through the joint covariance than an MCMC run performed only with the larger of the two subsets. The last column in Table 4 reports the significances for the second TPD-based estimator, comparing the differences of the joint and split TPDs directly, again calculated only for the $\boldsymbol{x}^D$ part of the data (compare also to the green/orange symbols in Fig. 8b). Those are all well below $1\,\sigma$ and based on this test we do not see any hints for tension in the KiDS-450 data either.

As the results of Efstathiou & Lemos (2018) are in general more comparable to the first TPD-based estimator, we suggest that their cross-validation approach is sensitive to the overall goodness of fit and does not directly indicate residual systematics in the data for a given split. This is further supported by the results from Troxel et al. (2018). As mentioned in Section 4.1, these authors propose to update the KiDS-450 covariance matrix with an improved shot-noise model, primarily incorporating previously neglected survey-boundary effects. Effectively, their proposed modifications increase the uncertainties in a scale-dependent manner, which relieves the tensions reported by Efstathiou & Lemos (2018) for all their data splits; very much in agreement with our TPD-based tension check.

6 CONCLUSIONS

We presented three tiers of Bayesian consistency checks for correlated data sets. These tests are based on a symmetric (as opposed to a cross-validation) approach in the sense of introducing independent parameter sets for each mutually exclusive split of the fiducial data set in the likelihood evaluations, while still linking them through their joint covariance accounting for the correlations between the (sub)data sets. In particular, these are used to calculate evidence ratios, i.e. Bayes factors, as the first tier of consistency checks and differences in inferred posterior parameters as the second tier. The third tier takes place in the data domain and for that we introduce the concept of TPDs, a special case of Bayesian PPDs.

We showcased the usage of the TPDs with analytically tractable toy models and gave an intuitive definition of the significance for tension based on the TPDs. Then, we proceeded to apply the consistency checks to real cosmic shear data from the KiDS-450 analysis by Hildebrandt et al. (2017) and re-assessed earlier systematics tests and claims of internal tensions.

The major conclusions of our analysis are as follows:

(i) There exist multiple well-posed definitions of tension significance, which asses different aspects of the data and the model. Here, we show that care needs to be taken in their interpretation and comparison with other results, as some of these methods are more sensitive to tensions within the different parts of the data (e.g. Fig. 8b), while the others quantify tension between the data and the model (e.g. Fig. 8a). As a consequence, an '$x\sigma$ tension' is not a universal statement.

(ii) The Bayes factor is only a necessary requirement that a comparison of data sets has to pass for consistency, but not a sufficient one (see also Raveri & Hu 2018 who arrive at a similar conclusion and Jenkins & Peacock 2011 for a general criticism of the Bayes factor as a reliable decision making tool). This is due to the prior volume which has a significant impact on the Bayes factor. Wide prior ranges – particularly on parameters that are only weakly constrained by the data – will lower the evidence in general. This can produce artificial consistency between inconsistent data sets; see for example fig. 10 in DES Collaboration (2017). Moreover, in our approach, the duplication of the full parameter space in the likelihood evaluation of the subsets lowers the evidence further. To mitigate both effects, one should only duplicate the key parameters that are constrained best by the data. As this complicates the implementation quite significantly, we leave the pursuit of this approach to future work.

(iii) The TPD-based consistency estimators are complementary to the Bayes factor and posterior space analyses by providing a means of finding the sources of tension in the data domain. Moreover, we can both quantify tension in the data and the goodness of fit of the model by comparing the TPDs derived from the joint and duplicated parameter set to each other or each individually to the data distribution (assumed to be multivariate Gaussian).

(iv) Applying the three tiers of consistency checks to the KiDS-450 tomographic cosmic shear correlation functions does not yield significant evidence for tension in any of the checks, contrary to previous claims in the literature. We find evidence that the reported significant tension was driven by not fully accounting for the strong correlations in the data across splits, by a stricter definition of tension significance, and by an approach that mixes overall model fit quality with actual tension between the data splits. Indeed, an improved data covariance model was recently reported to alleviate the previously claimed tension to negligible levels (Troxel et al. 2018), in line with the results for our TPD-based tension estimate on the *original* KiDS-450 data set. The impact of improved modelling, including the covariance, on the internal consistency of KiDS weak-lensing data is investigated in Hildebrandt et al. (2018).

The core calculations for all our consistency checks are based on performing joint likelihood evaluations for mutually exclusive subsets still linked through the joint covariance but separated in terms of parameter sets and parameter-dependent calculations. For that purpose, we modified the likelihood evaluation code MONTE PYTHON and this modified version (and the likelihoods) are made publicly available.[7] As long as the likelihood analysis is performed with an algorithm that readily produces the evidence (such as nested sampling), the main computational cost of our consistency tests lies in the doubling of the parameter space to be sampled. For the current analysis choices, this is readily tackled by MULTINEST, while for the increased nuisance parameter spaces expected for forthcoming studies, it may be advisable to limit the duplication to cosmological and/or astrophysical parameters.

Since our tests are by design sensitive to any inconsistencies in the data, it may be challenging to integrate them into blinded analyses. Great care has to be taken that the blinding procedure preserves consistency within the data set, and particularly also across all probes to be combined. Consistency checks of the kind presented in this work are always conditional on the model that is fitted and as such necessarily involve the computation of parameter posteriors, which may be prohibited in strict implementations of blinding until the very final stages of the analysis. We consider it acceptable to run the consistency tests after unblinding; however, it is then paramount to fix the choice of data splits beforehand.

Finally, we emphasize again that the consistency checks demonstrated here on cosmic shear data are fully general and can be applied to any (correlated) data set for which one can evaluate its likelihood function and approximate it as multivariate normal.

[7]Modified '2cosmos' MONTE PYTHON (including the corresponding likelihood): https://github.com/fkoehlin/montepython_2cosmos_public Likelihood for KiDS-450 data to be used within standard MONTE PYTHON: https://github.com/fkoehlin/kids450_cf_likelihood_public

In that regard, the consistency checks can also prove to be very useful for establishing the internal consistency of each probe used in multiprobe analyses such as were carried out for KiDS (van Uitert et al. 2018; Joudaki et al. 2018) and DES (DES Collaboration 2017) already. In the near future, these surveys will be surpassed by even bigger large-scale structure surveys such as those carried out by the spaceborne *Euclid* (Laureijs et al. 2011) and *WFIRST*[8] satellites or the ground-based DESI (Levi et al. 2013) and LSST (Ivezic et al. 2008). We anticipate consistency tests like the ones presented in this work to become an integral part of the analysis pipelines within these surveys, and instrumental for the joint cosmological inference across probes.

ACKNOWLEDGEMENTS

We thank H. Peiris, G. Efstathiou, and the participants of the Understanding Cosmological Observations meeting at the Centro de Ciencias de Benasque for insightful discussions. We would also like to thank K. Kuijken for comments, H. Hildebrandt for testing (parts of) this methodology and pipeline, and H. Hoekstra for computational resources. We also appreciate the very helpful and constructive comments of the anonymous referee which helped to further improve the presentation of this work. FK acknowledges support from the World Premier International Research Center Initiative (WPI), MEXT, Japan and from JSPS KAKENHI grant number JP17H06599. BJ acknowledges support by the UCL Cosmoparticle Initiative. MA acknowledges support from the ERC under grant agreement 647112. SJ acknowledges support from the Beecroft Trust and ERC 693024. TT acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 797794.

Based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017, and 177.A-3018, and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD, and the KiDS production team, on behalf of 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 (FK, BJ, MA, and MV), followed by one alphabetical group. The alphabetical group (SJ and TT) includes those who have either made a significant contribution to the data products, or to the scientific analysis.

REFERENCES

Adhikari S., Huterer D., 2019, J. Cosmol. Astropart. Phys., 2019, 036
Audren B., Lesgourgues J., 2011, J. Cosmol. Astropart. Phys., 2011, 037
Audren B., Lesgourgues J., Benabed K., Prunet S., 2013, J. Cosmol. Astropart. Phys., 2013, 001
Bartelmann M., Schneider P., 2001, Phys. Rep., 340, 291
Begeman K., Belikov A. N., Boxhoorn D. R., Valentijn E. A., 2013, Exp. Astron., 35, 1
Benítez N., 2000, ApJ, 536, 571
Blas D., Lesgourgues J., Tram T., 2011, J. Cosmol. Astropart. Phys., 2011, 034
Bridle S., King L., 2007, New J. Phys., 9, 444
Buchner J. et al., 2014, A&A, 564, A125
Charnock T., Battye R. A., Moss A., 2017, Phys. Rev. D, 95, 123535
de Jong J. T. A. et al., 2015, A&A, 582, A62
DES Collaboration, 2017, preprint (arXiv:1708.01530)
Efstathiou G., Lemos P., 2018, MNRAS, 476, 151
Erben T. et al., 2013, MNRAS, 433, 2545
Feeney S. M., Peiris H. V., Williamson A. R., Nissanke S. M., Mortlock D. J., Alsing J., Scolnic D., 2018, preprint (arXiv:1802.03404)
Fenech Conti I., Herbonnet R., Hoekstra H., Merten J., Miller L., Viola M., 2017, MNRAS, 467, 1627
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2013, preprint (arXiv:1306.2144)
Gelman A., Stern H. S., Carlin J. B., Dunson D. B., Vehtari A., Rubin D. B., 2013, in Bayesian data analysis. Chapman and Hall/CRC, UK
Harnois-Déraps J., Waerbeke L. v., Viola M., Heymans C., 2015, MNRAS, 450, 1212
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hildebrandt H. et al., 2018, preprint (arXiv:1812.06076)
Hirata C. M., Seljak U., 2004, Phys. Rev. D, 70, 063526
Ivezic Z. et al., 2008, preprint (arXiv:0805.2366)
Jeffreys K., 1961, in Theory of probability, Oxford University Press, UK
Jenkins C. R., Peacock J. A., 2011, MNRAS, 413, 2895
Joachimi B., Mandelbaum R., Abdalla F. B., Bridle S. L., 2011, A&A, 527, 26
Joudaki S. et al., 2017, MNRAS, 471, 1259
Joudaki S. et al., 2018, MNRAS, 474, 4894
Kaiser N., 1992, ApJ, 388, 272
Kilbinger M., 2015, Rep. Prog. Phys., 78, 086901
Köhlinger F. et al., 2017, MNRAS, 471, 4412
Kuijken K. et al., 2015, MNRAS, 454, 3500
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Levi M. et al., 2013, preprint (arXiv:1308.0847)
Limber D. N., 1953, ApJ, 117, 134
Lin W., Ishak M., 2017a, Phys. Rev. D, 96, 023532
Lin W., Ishak M., 2017b, Phys. Rev. D, 96, 083532
LoVerde M., Afshordi N., 2008, Phys. Rev. D, 78, 123506
Mandelbaum R. et al., 2018, PASJ, 70, S25
Marshall P., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302
Mead A. J., Peacock J. A., Heymans C., Joudaki S., Heavens A. F., 2015, MNRAS, 454, 1958
Mead A. J., Heymans C., Lombriser L., Peacock J. A., Steele O. I., Winther H. A., 2016, MNRAS, 459, 1468
Miller L. et al., 2013, MNRAS, 429, 2858
Planck Collaboration XIII, 2016, A&A, 594, A13
Planck Collaboration LI, 2017, A&A, 607, A95
Planck Collaboration VI, 2018, preprint (arXiv:1807.06209)
Raveri M., Hu W., 2018, preprint (arXiv:1806.04649)
Riess A. G. et al., 2016, ApJ, 826, 56
Riess A. G. et al., 2018, ApJ, 861, 126
Schaye J. et al., 2010, MNRAS, 402, 1536
Scott D. W., 1992, Multivariate Density Estimation. Wiley, New York
Semboloni E., Hoekstra H., Schaye J., van Daalen M. P., McCarthy I. G., 2011, MNRAS, 417, 2020
Semboloni E., Hoekstra H., Schaye J., 2013, MNRAS, 434, 148
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Troxel M. A. et al., 2018, MNRAS, 479, 4998
van Daalen M. P., Schaye J., Booth C. M., Dalla Vecchia C., 2011, MNRAS, 415, 3649
van Uitert E. et al., 2018, MNRAS, 476, 4662

APPENDIX A: SENSITIVITY ANALYSIS

Here, we present a sensitivity analysis for two key types of systematics that we would like to be able to detect in a data vector of a cosmic shear survey:

[8] wfirst.gsfc.nasa.gov

(i) a shift in the mean of the source redshift distribution for any of the redshift bins and

(ii) the effect of a systematic generating B modes.

These two particular systematics are motivated by the findings of Hildebrandt et al. (2017), Joudaki et al. (2017), van Uitert et al. (2018), and Efstathiou & Lemos (2018) in the KiDS-450 data; see Section 5 for details.

To test the sensitivity of the estimators for these two systematics, we first create a noise-free mock $\xi_\pm$ data vector based on an arbitrary sampled set of parameters from KiDS-450 cosmologies derived with our likelihood pipeline.[9] In a second step, we perturb this fiducial noise-free mock vector to include the effects of the two systematics as follows: the shift in the source redshift distribution is created through a mock data vector obtained from the true $n(z)$ at a shifted redshift $z + \mathrm{d}z$:

$$\tilde{n}(z) = n(z + \mathrm{d}z), \tag{A1}$$

i.e. a shift by positive $\mathrm{d}z$ moves the distribution $\tilde{n}(z)$ to lower redshifts. For the B-mode systematic, we add a fraction f_B of the real B modes measured in the KiDS-450 data (cf. appendix D6 of Hildebrandt et al. 2017) to the ξ_+ part of the fiducial mock data vector,

$$\xi_+(\theta) = \xi_+^\mathrm{mocks} + 2\, f_\mathrm{B}\, \xi_\mathrm{B}(\theta), \tag{A2}$$

which assumes that the systematic adds equally to the E- and B-mode channels.[10]

A1 Sensitivity of the Bayes factor

Following Section 2.1, we perform the sensitivity analysis as follows: we use the fiducial $\xi_\pm$ mock data vector and the covariance matrix of KiDS-450 for the fiducial scales as the input data for the joint MCMC run (i.e. the numerator of equation 3) corresponding to the model H_0: 'there exists one common set of parameters that describe all data sets' and sample the likelihood in the same parameters and prior ranges as presented in Hildebrandt et al. (2017) with the caveats discussed in Section 4.1.

For the split MCMC run (i.e. the denominator of equation 3) which tests now the model H_1: 'there exist two separate parameter sets that each describe one subset of the data', we split the fiducial mock data vector according to the systematic we want to test. For example, for the shift in the source redshift distribution we split the mock data vector $\boldsymbol{d}_\mathrm{tot}^\mathrm{mock}$ into one set $\boldsymbol{d}_\mathrm{a}^\mathrm{mock}$ containing the perturbed z-bin (and all its CCs) and the mutually exclusive set $\boldsymbol{d}_\mathrm{b}^\mathrm{mock}$ containing all other unperturbed z-bins (and their CCs), thus $\boldsymbol{d}_\mathrm{tot}^{\mathrm{mock}^\tau} = \{\boldsymbol{d}_\mathrm{a}^{\mathrm{mock}^\tau}, \boldsymbol{d}_\mathrm{b}^{\mathrm{mock}^\tau}\}$. In the case of adding a fraction of B modes, we split the data vector into its ξ_+ and ξ_- parts, as the assumed systematic only contributes to ξ_+.

We estimate the evidences for every joint and split MCMC run for every increment (in $\mathrm{d}z$ or f_B) of the systematic in question, i.e. the joint and split runs are fitted to increasingly perturbed mock data vectors. The perturbation is not taken into account in the model. That way we calculate the evidence ratio as a function of increasing deviation from the unperturbed mock data vector.

The sensitivity of the Bayes factor to increasing systematic shifts is presented in Fig. A1. In the left-hand panel (Fig. A1a), we shift the source redshift distribution of z-bin 3. We note that the choice between the four fiducial z-bins is not entirely arbitrary, since the lower z-bins have lower S/N than the higher z-bins. As shown in that panel the Bayes factor only starts to flag fairly large shifts of $\mathrm{d}z > 0.15$ as problematic (i.e. at least finding 'substantial' evidence on Jeffreys' scale for the alternative model H_1). Further we note that the slope in Fig. A1(a) is another sign that the Bayes factor is not well suited for quantifying tension between the splits of the data set, as all classifications on Jeffreys' scale occur within a tiny span of $\Delta\mathrm{d}z \sim 0.02$.

In the right-hand panel (Fig. A1b), we show the sensitivity of the Bayes factor to adding a fraction of B modes, f_B, to the fiducial mock data vector mimicking the effect of a systematic adding equal power in E and B modes. Only fractions of $f_\mathrm{B} > 1.60$ of measured KiDS-450 B modes added to ξ_+ are flagged as problematic by this test.

The weak sensitivity of the Bayes factor test is not completely unexpected: first, doubling the parameter space in every split MCMC run is a conservative approach because Occam's razor is integral to the evidence calculation. Thus, a model with a significantly increased parameter space is strongly disfavoured a priori (if it does not provide a significantly better fit to the data). Secondly, the prior ranges are also entering in the evidence calculation and any (unphysically) wide prior range (as is the case in our example, e.g. for the prior on $\Omega_\mathrm{cdm}\, h^2$)[11] will also decrease the evidence for the model being tested, again severely disfavouring the split model (unless it explains the data significantly better, see also Raveri & Hu 2018).

A2 Sensitivity of duplicate parameter differences

In addition to the Bayes factor, we also look at how the two systematics affect the inferred posteriors directly. Following Section 2.2, the comparison of the (key) parameters obtained for each of the two mutually exclusive subsets of the noise-free data vector used in the split MCMC run is straightforward to interpret: if the duplication of the parameters is indeed unnecessary (i.e. there is no tension in the data set) both subsets (which are still coupled through the full covariance by construction) should produce close-to identical posteriors.

In Fig. A2, we show the differences in 2D projections of key parameters obtained from each subset of the split MCMC runs for both systematic tests. The left-hand panel (Fig. A2a) shows the results for a shifted source redshift distribution of z-bin 3 and the right-hand panel (Fig. A2b), the same for the added fraction of small-scale B modes measured in KiDS-450. In both panels, the dashed contours indicate the differences of parameters for the unperturbed noise-free mock data vector at 68 per cent and 95 per cent credibility. The solid contours show the differences of parameters for the strongest amplitude of each systematic in the test (i.e. $\mathrm{d}z_3 = 0.30$ for the shift in the source redshift distribution of z-bin 3 and $f_\mathrm{B} = 2.00$ for the added small-scale B modes). The dotted line connecting the centres of the contours in each subpanel

[9] $\Omega_\mathrm{cdm}\, h^2 = 0.1014$, $\Omega_\mathrm{b}\, h^2 = 0.0199$, $\ln(10^{10} A_\mathrm{s}) = 3.3013$, $h = 0.7702$, $n_\mathrm{s} = 1.2256$, $A_\mathrm{IA} = 1.7861$, and $A_\mathrm{bary} = 1.8681$; yielding $\Omega_\mathrm{m} = 0.2044$, $\sigma_8 = 0.9837$, and $S_8 = 0.8119$.

[10] The factor of 2 on the right-hand side of equation (A2) arises because we assume the systematic contribution to the ellipticity measurement is uncorrelated with the true sheared ellipticity $\varepsilon_\mathrm{true}$, and that it adds linearly such that $\varepsilon_\mathrm{obs} = \varepsilon_\mathrm{true} + \varepsilon_\mathrm{sys}$. For a detailed derivation, we refer the reader to appendix D6 in Hildebrandt et al. (2017).

[11] The prior range, however, was intentionally chosen in Hildebrandt et al. (2017) to be that wide in order to guarantee a full sampling of the Ω_m versus σ_8 degeneracy plane.

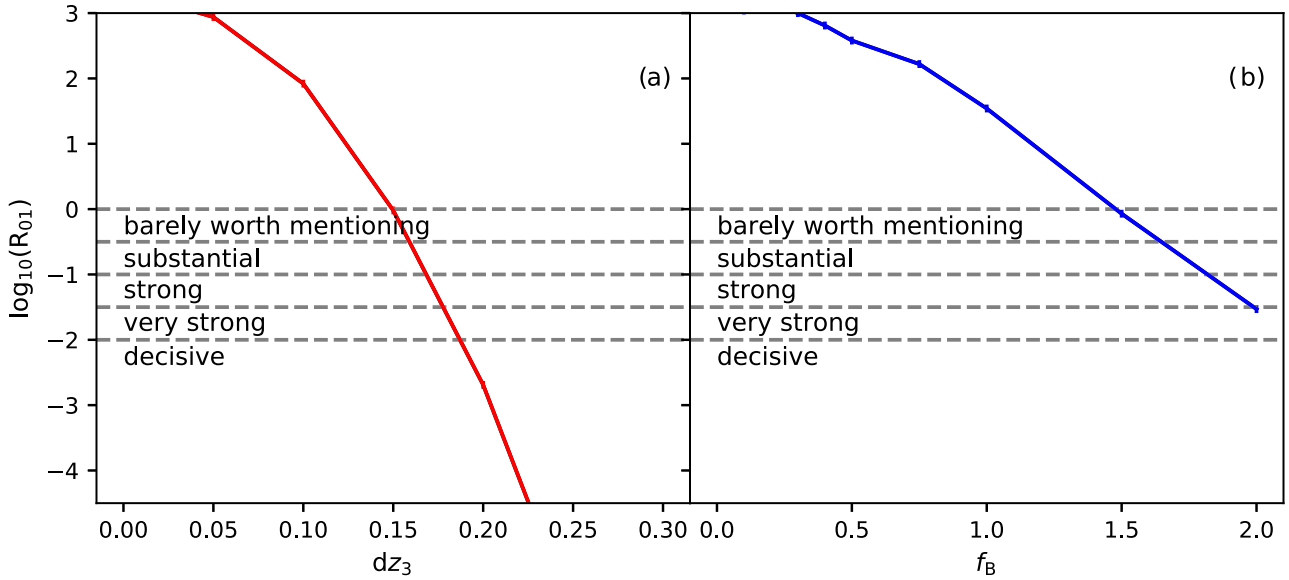

Figure A1. The common logarithm of the evidence ratio R_{01} as a function of (a) an additional shift dz of the source redshift distribution for z-bin 3. (b) Adding a fraction f_B of small-scale B modes measured in KiDS-450 to ξ_+ (see the text for details). Both systematics are added on top a mock data vector mimicking the KiDS-450 correlation function vector. In both panels, we interpret the Bayes factor in terms of Jeffreys' scale and the statements should be read as 'barely worth mentioning', 'substantial', etc., evidence for H_1 : 'there exist two separate parameter sets that each describe one subset of the data'.

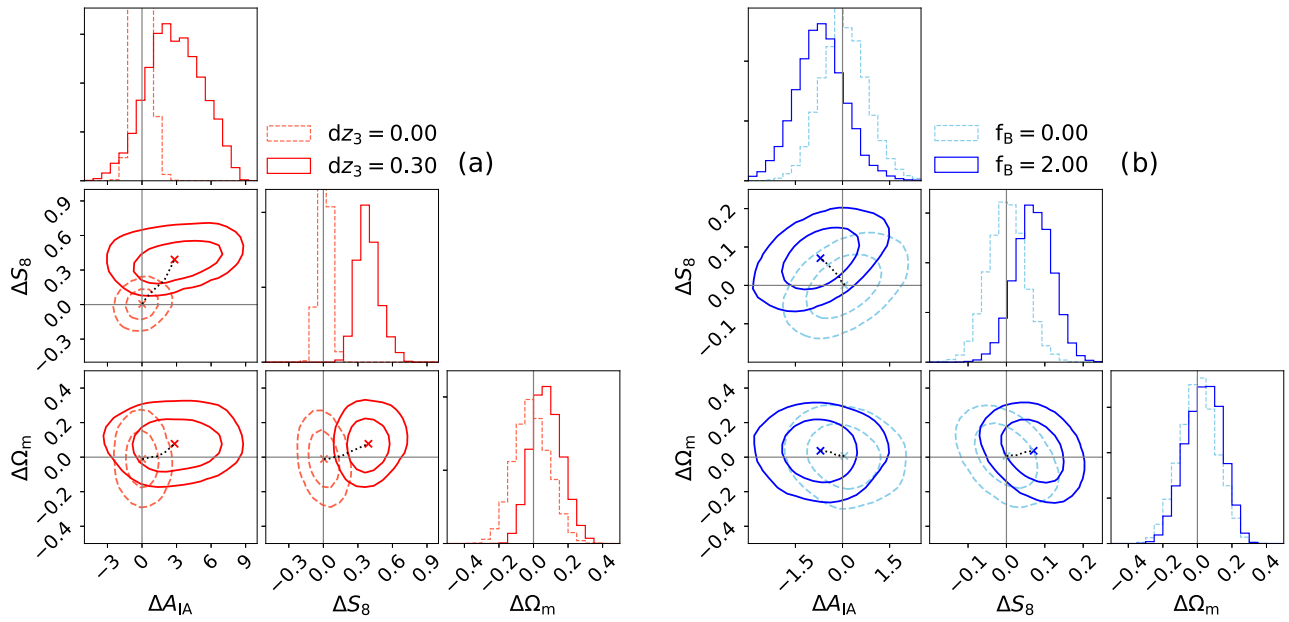

Figure A2. Duplicate parameter differences from the split MCMC run for 2D projections of key parameters for both sensitivity tests: (a) shifts of the source redshift distribution of z-bin 3 by d$z = 0$ (dashed contours) and d$z = 0.30$ (solid contours). The split applied in the duplicate parameter MCMC run is thus 'z-bin 3 (and all its CC) versus all other z-bin combinations', (b) adding fractions of $f_B = 0$ (dashed contours) and $f_B = 2.00$ (solid contours) of measured small-scale B modes in KiDS-450 to ξ_+ (see the text for details). The split applied in the duplicate parameter MCMC run is thus 'ξ_+ versus ξ_-'. The covariance between the mutually exclusive subsets of the split MCMC run is fully taken into account for the parameter inference. The dotted black lines in both panels correspond to the differences in the weighted means of the parameters of interest for all intermediate systematic shifts (cf. Fig. A1).

marks the position of the mean deviation as a function of increasing strength of the systematic.

As expected, the contours for the unperturbed mock data vector (dashed) are centred on zero, i.e. both splits of the data vector produce very similar posteriors. In contrast, most of the solid contours show displacements from zero revealing biases in the various 2D parameter projections. Especially the comparison of the dashed and solid contours in the 2D projections of S_8 or Ω_m versus A_{IA} in Fig. A2(a) is particularly interesting: in addition to introducing significant biases for all of the parameters, a part of the effect of shifting z-bin 3 (and all its CCs) is absorbed by the intrinsic alignment parameter A_{IA} whose contours broaden

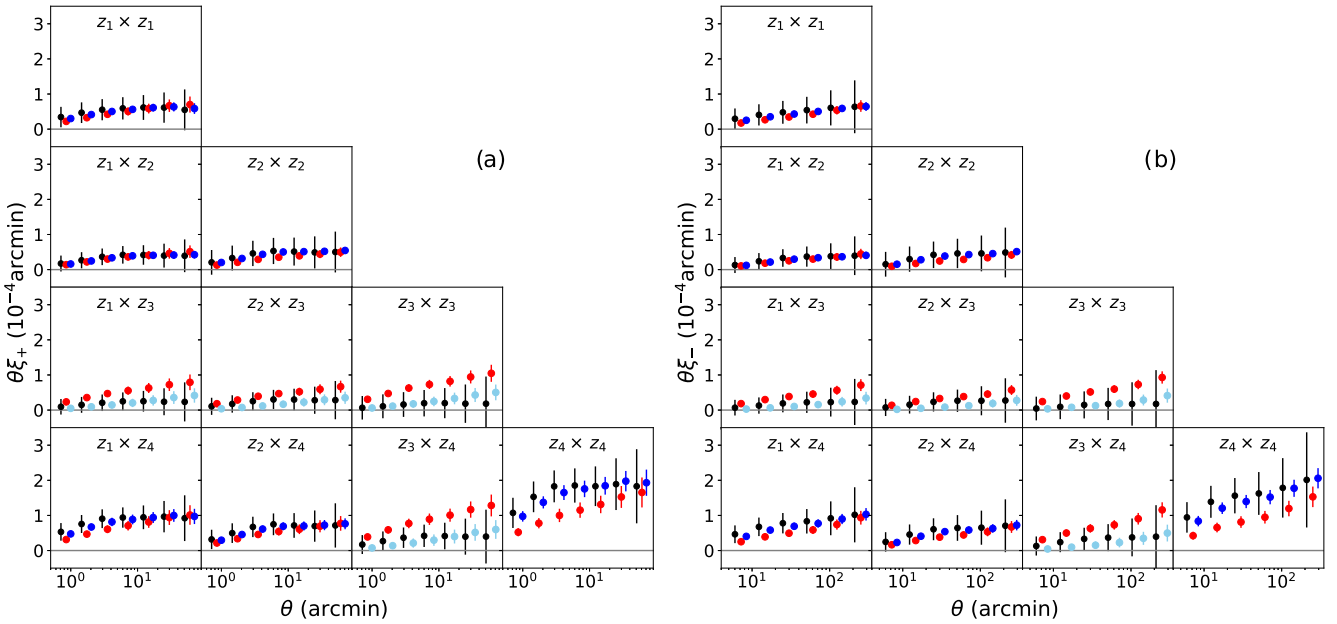

Figure A3. (a) Means of the TPD for ξ_+ correlation functions per angular scale, θ, and per redshift bin combination, $z_i \times z_j$. The TPDs are based on the joint and split cosmological and nuisance parameters (red and light/dark blue points) from MCMC runs fit to a noise-free mock data vector (black points) whose source redshift distribution was shifted by $dz = 0.30$ in z-bin 3. The split in both panels corresponds to 'z-bin 3 (and all its CCs)' (light blue points) versus 'all other redshift bin combinations' (dark blue points). Error bars are derived from the 68 per cent credibility interval around the mean. The error bars for the KiDS-450 data points are based on the diagonal of the fiducial covariance matrix. (b) The same as in (a), but for the ξ_- estimator.

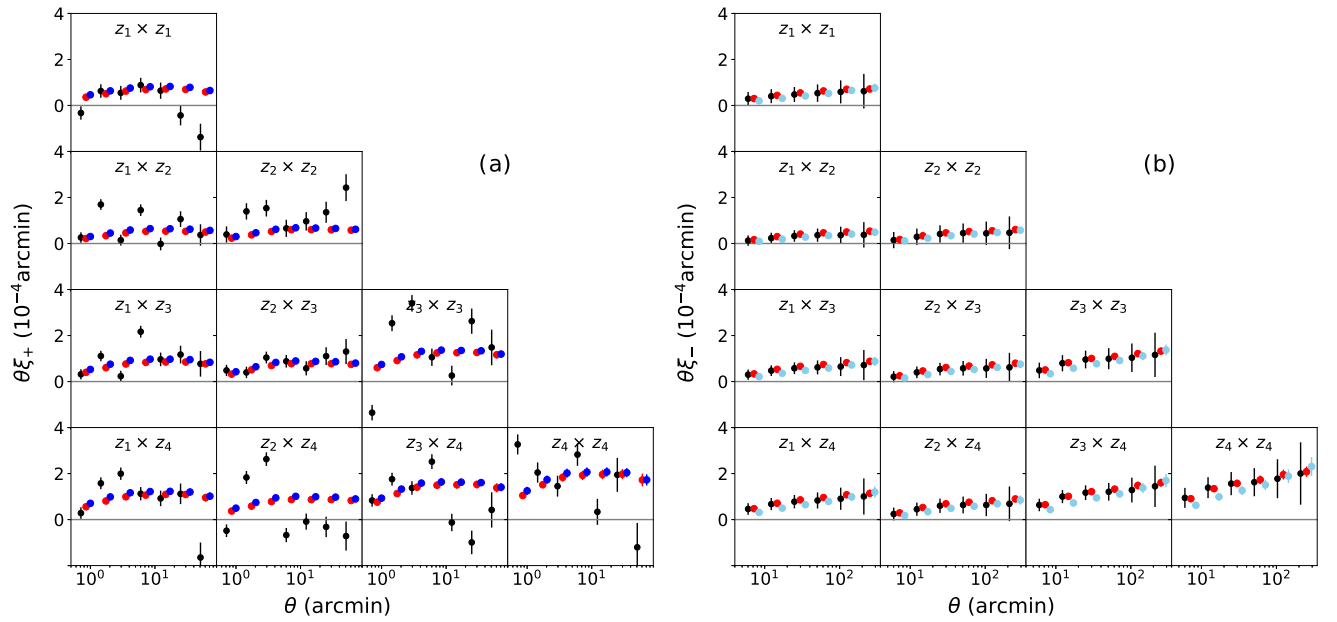

Figure A4. (a) Means of the TPD for the ξ_+ estimator based on the joint and split cosmological and nuisance parameters (red and blue points) compared and fit to a noise-free mock data vector (black points) to which a fraction of $f_B = 2.00$ of measured B modes were added to ξ_+. The split in both panels corresponds to 'ξ_+' (dark blue points) versus 'ξ_-' (light blue points, not visible). Error bars on the means are derived from the 68 per cent credibility interval around the mean. The error bars for the data are based on the diagonal of the covariance matrix. (b) The same as in (a) but for the tomographic ξ_- correlation functions.

significantly for the strongest systematic shift. This implies that the intrinsic alignment amplitude can absorb the effect of a bias in the redshift source distributions (in part), and broad errors on A_{IA} are a typical signature for that taking place.

Quantifying the tension over the three-parameter subspace of ΔS_8, $\Delta\Omega_m$, and ΔA_{IA} as outlined in Section 2.2 yields no tension for the unperturbed mock data vector for both systematics (i.e. the dashed contours in Fig. A2). For the most extreme shift in the sensitivity analysis of $dz_3 = 0.30$ in the source redshift distribution of z-bin 3, we find a significant tension of $5.92\,\sigma$. For the most extreme case of the added B-mode systematic with $f_B = 2.00$ the tension is at $2.03\,\sigma$ over the full three-parameter subspace.

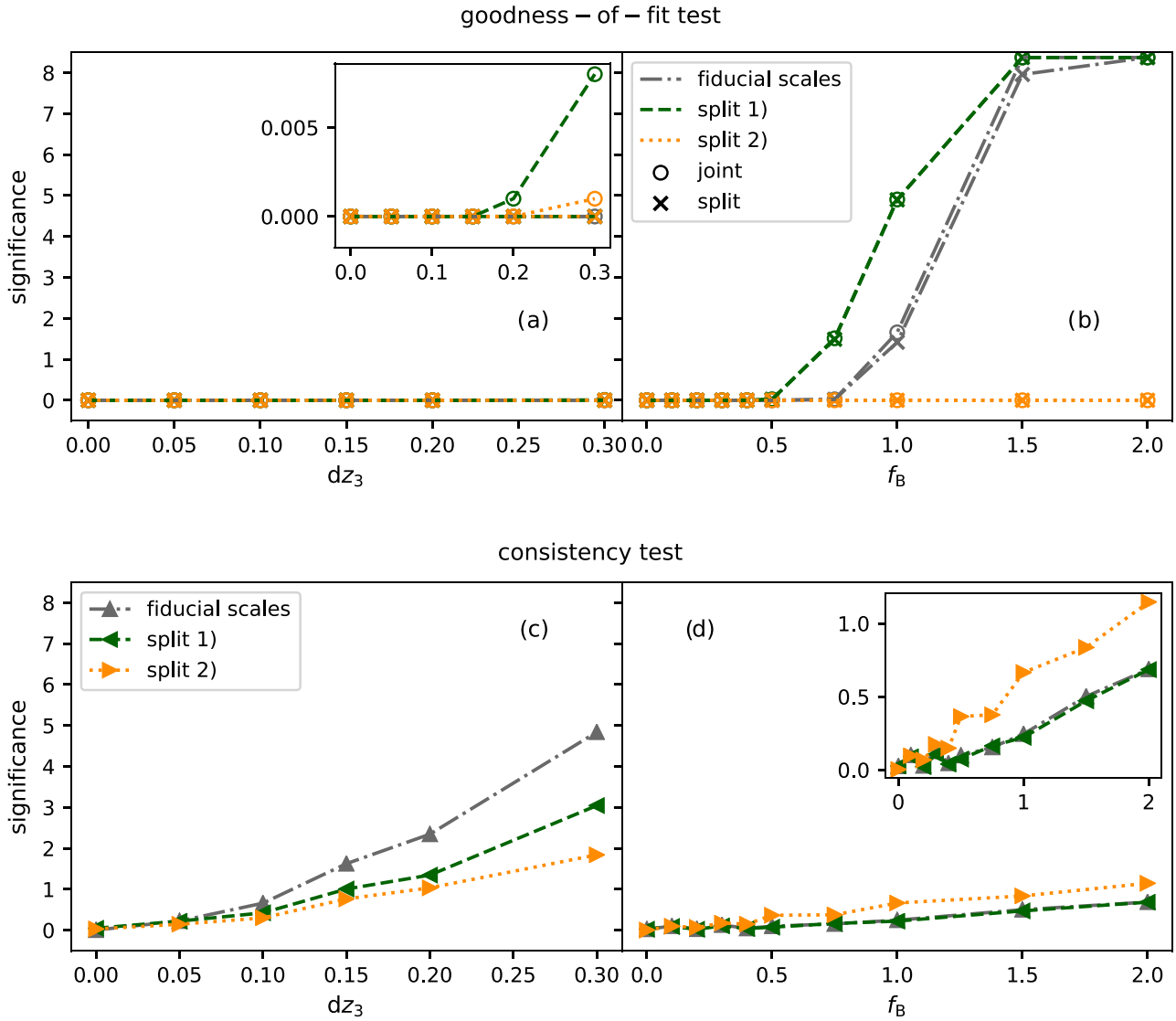

Figure A5. Significance as derived with the two TPD-based consistency estimators for a noise-free mock data vector and as a function of two increasing systematics at a time. (a) The significances are derived by comparing the mock data and TPD distributions as a function of the first key systematic, i.e. a shift $\mathrm{d}z$ applied to the source redshift distribution of z-bin 3. Lines with circles only use the joint MCMC for the TPDs, whereas lines with crosses are based on the split MCMC. The different line styles and colours indicate which selection was applied in the calculation of the significance: the dashed–dotted lines (grey) use all fiducial scales (Table 1), the dashed lines (green) use only 'split 1' corresponding to z-bin 3 (and all its CCs) and the dotted lines (orange) 'split 2', i.e. all other z-bin combinations. Note that all lines overlap along the zero line. Therefore, we also provide a zoom-in panel. (b) The same as in (a) but as a function of the second key systematic, i.e. adding a fraction f_B of the measured B modes in KiDS-450 to the ξ_+ part of the mock data vector. Hence, the dashed lines (green) correspond to the ξ_+ mask and the dotted lines (orange) correspond to the ξ_- mask. (c) The significances are now derived by comparing the differences between the joint and the split TPDs to zero for the first key systematic as in (a). (d) The same as in (c) but for the second key systematic as in (b).

A3 Sensitivity of translated posterior distributions

We now turn towards comparisons in the data domain and thus to the TPDs introduced in Section 2.3, characterizing their behaviour for the two systematics. For the calculation of the TPDs, we approximate the integration of equation (5) by using the converged MCMCs from the evidence and Bayes factor calculations. For each sample of cosmological and nuisance parameters in the chain, we recreate the $\xi_\pm$ theory vector and the distribution of these represents the TPD. To guide the reader's intuition for the interpretation of the TPDs and to also demonstrate their strength as a qualitative diagnostics tool, we show in Figs A3(a) and (b) the means of the ξ_+ and ξ_- estimators for the split and joint TPDs created from the cosmology and nuisance parameters fit to a mock data vector whose source redshift distribution was shifted by $\mathrm{d}z = 0.30$ for z-bin 3.

For a first qualitative interpretation, we compare the joint and split TPDs to the data vector and to each other: the split TPD follows the mock data vector very closely in all subpanels of the triangle that contain z-bin 3 (light blue points) and also yields a reasonable fit in all other subpanels (dark blue points). This is expected since the MCMC from which the split TPDs are derived have one parameter set for the z-bin 3 (and all its CCs) part of the mock data vector and one additional set for the remaining part of the vector. The posteriors, however, are still linked through the full covariance of the data set and this is also visible in the TPDs. For example, the

dark blue points are biased low with respect to the mock data vector in the autocorrelation panel of z-bin 4. This is due to the shift of z-bin 3 (to lower values) and the corresponding close match of the cyan points propagating to z-bin 4 because of the strong correlations between different redshift bin combinations (cf. Appendix C for the impact of correlations in the data in an analytical test case). The behaviour of the joint TPD, however, is quite different. Since it is derived from the MCMC with only one parameter set for the full mock data vector, the joint TPD shows large deviations with respect to the shifted mock data vector in all subpanels containing z-bin 3 and a reasonable match to the more constraining mock data in all other subpanels.

In Fig. A4, we show the corresponding plots for the added B-mode systematic. In the left-hand panel (Fig. A4a), we demonstrate the largest impact of this effect considered in our analysis by adding a fraction $f_{\rm B} = 2.00$ of the small-scale B modes measured in KiDS-450 to the ξ_+ estimator (cf. equation A2), whereas the right-hand panel shows the (unchanged) ξ_- estimator for that case. In both panels the mock data vector is shown in black with error bars derived from the fiducial KiDS-450 covariance matrix. The means of the TPD based on the joint MCMC are given as red points and the means of the two TPDs derived from the split MCMC are shown in light and dark blue. Due to splitting the data set into its ξ_+ and ξ_- parts, Fig. A4(a) contains only dark blue and consequently Fig. A4(b) only light blue points.

The theoretical model is not able to capture the distortions due to adding B modes as shown in Fig. A4(a) and the difference between the joint and split TPDs is negligible in this panel. This is in contrast to the behaviour of the TPDs for the z-bin shift systematic. There, the split TPD tracing the shifted part of the mock data vector yields a close match to the perturbed mock data, whereas the joint TPD shows larger deviations in these panels due to that the overall fit is dominated by the other unperturbed part of the mock data. Hence, this indicates again that comparing the individual joint and split TPDs to the data distribution is a quantification of the overall goodness of fit, whereas a comparison of the joint to the split TPDs is sensitive to the tension between the splits.

To make the previous discussion quantitative, we apply the TPD-based consistency estimators defined in Section 2.3 to the two systematics and show the results in the panels of Fig. A5. Both upper and lower left panels refer to the results for an increasing shift, $\mathrm{d}z_3$, in the source redshift distribution of z-bin 3, whereas the upper and lower right panels depict the case for adding an increasing fraction $f_{\rm B}$ of B modes to the ξ_+ estimator.

In the upper left and right panels, we compare the TPDs from the joint and split MCMC runs individually to the (mock) data distribution as described in Section 2.3, i.e. the reported significances, σ, correspond to the highest σ-value for which $\mathrm{I}_{\rm TPD} \leq 1 - c_{\rm m}$ as a function of increasing systematic shift. In these panels, the circle and cross symbols denote whether the results are derived from the joint or split MCMC runs, respectively. In addition to that the different colours and line styles correspond to the selection applied to the TPD vectors: dashed–dotted lines (grey) indicate that the fiducial scales of the Hildebrandt et al. (2017) analysis (see also Table 1) were applied to the selection of the TPD and mock data vectors in the comparison. The dashed (green) and dotted (orange) lines correspond to applying the selection of the splits, i.e. 'split 1' corresponds to z-bin 3 (and all its CCs) and 'split 2' to all other z-bin correlations in Fig. A5(a). In Fig. A5(b), 'split 1' corresponds to ξ_+ and 'split 2' to ξ_-.

Focusing at first on the right-hand panel, we observe that only the ξ_+ subset (green dashed lines) shows a rise in the significance for an increasing fraction of added B modes, $f_{\rm B}$. In contrast to that the significance for the ξ_- subset remains constant around zero (orange dotted lines). Moreover, there is (almost) no difference between comparing the joint (circles) or split TPDs (crosses) to the mock data. This is a consequence of that estimator measuring the general goodness of fit of the model rather than the internal tension in the data. This interpretation is further supported by comparing TPDs and mock data for the fiducial data vector (grey dashed–dotted lines) as this reproduces features almost identical in the significance to the ξ_+ subset (green dashed lines). We also note that the tension significances are higher compared to those from the Bayes factor for this systematic (see Fig. A1b).

Moving on to the z-bin shift systematic, we note the generally low level of significance estimated in all cases. Zooming-in, however, reveals some interesting and expected features such as an increase in significance for an increasing shift of z-bin 3, whereas the complementary subset containing 'all other' z-bin combinations (orange dotted lines) remains constant around zero. Moreover, the increase in significance is only a feature for the TPDs derived from the joint MCMC (circles), whereas the significances derived from the split TPDs (crosses) remain flat. For comparison, we also show the significance for tension when applying the fiducial scales selection to the mock data and TPDs (grey dashed–dotted lines; see Table 1) which do not produce any trends at all.

To summarize, the comparison between the data distribution and the individual (split or joint) TPDs does *not* quantify tension but the overall goodness of fit of the model. A likely scenario for this is that there is a systematic effect present in the data, but the chosen split of the data vector has failed to separate elements with significant contamination by the systematic from those with no or only small contributions.

Therefore, we now consider the *differences* between the joint and split TPDs instead of comparing each TPD individually to the (mock) data distribution: if there is no tension in the (mock) data, we expect the difference of the TPDs to be consistent with zero. If the split succeeds in isolating the data affected by a systematic, a significant discrepancy between the joint and split TPDs is expected. For this approach, we also propagate the correlations between the joint and split TPDs into the uncertainty of their difference distribution as both distributions are derived from the same data. Another complication arising in the quantification of tension for this approach is that we are predicting a typically $\sim$100D distribution from a $\sim$10D posterior distribution, so the estimated covariance of the former will necessarily be close to being fully correlated. Therefore, we need to employ a PCA on the covariance matrix keeping only the components that contain at least 95 per cent of the variance for the inversion of the matrix. The inverse is required in the fitting-to-zero procedure based on which we assign significances. For the full details on this, we refer the reader to Appendix B. We show the resulting significances in the two lower panels of Fig. A5.

The level of the significance in the left-hand panel for the z-bin shift systematic has increased to values in accordance with our expectation from the Bayes factor analysis (cf. Fig. A1a). The z-bin 3 (and all its CCs) subset produces now tension at the $\sim 3\,\sigma$ level for the largest shift of $\mathrm{d}z = 0.30$. Moreover, the significances in the corresponding panel for the added B-mode fractions, i.e. Fig. A5(d), have decreased in accordance with our expectations from the Bayes factor analysis (cf. Fig. A1b). However, the roles of ξ_+ and ξ_- seem to be reversed in these panels: since only the ξ_+ part is directly modified by adding the B-mode fraction $f_{\rm B}$,

we would have naively expected to find higher significances when comparing the joint to split TPDs for the ξ_+ correlation functions because the ξ_- correlation functions can only be affected by the B modes through the joint covariance of both splits. However, we know from the previous TPD-based estimator that the theory vector is in general not a good fit to the ξ_+ correlation functions. Therefore, the difference of the TPDs is much more sensitive to differences in the ξ_- correlation functions for which the theory vector provides a better match to the mock data.

APPENDIX B: ERROR ESTIMATION AND CORRELATION PROPAGATION FOR THE TPD-BASED TENSION ESTIMATOR

The second TPD-based consistency estimator introduced in Section 2.3 compares the differences between the joint and split TPDs to zero in order to assign a significance for consistency/tension. In practice, the joint and split TPDs are derived in independent calculations and we set the magnitude of the error bars of the differences by adding the diagonals of the covariances directly estimated from the joint and split TPDs (from on the order of 10^4 samples each). We know, however, that these errors are correlated since each 'independent' MCMC run uses the same data. Therefore, we need to propagate these correlations into the final uncertainties of the TPD of the differences and the significances derived from those. For that, we employ a Fisher matrix analysis and start with writing out the total data vector as

$$\boldsymbol{d}^{\tau}_{\rm tot} = \{\boldsymbol{d}^{\tau}_{\rm a}, \boldsymbol{d}^{\tau}_{\rm b}\}, \tag{B1}$$

where $\boldsymbol{d}_{\rm a}$ and $\boldsymbol{d}_{\rm b}$ correspond to the mutually exclusive splits with S and $N-S$ entries, respectively, for in total N entries in $\boldsymbol{d}_{\rm tot}$. Based on this data vector, we want to derive a parameter set

$$\boldsymbol{p}^{\tau}_{\rm tot} = \{\boldsymbol{p}^{\tau}_{\rm j}, \boldsymbol{p}^{\tau}_{\rm s_a}, \boldsymbol{p}^{\tau}_{\rm s_b}\}, \tag{B2}$$

where the index j labels the parameter set from the joint MCMC and the indices $s_{\rm a}$ and $s_{\rm b}$ label the ones derived from the subsets of the split MCMC. With these definitions, we can now write down the Fisher information matrix $\mathbf{F}$:

$$(\mathbf{F})_{\mu\nu} = \sum_{i,j}^{N} \frac{\partial \boldsymbol{d}_{\rm tot,i}}{\partial \boldsymbol{p}_{\rm tot,\mu}} (\mathbf{C}^{-1}_{\rm data})_{\rm ij} \frac{\partial \boldsymbol{d}_{\rm tot,j}}{\partial \boldsymbol{p}_{\rm tot,\nu}}. \tag{B3}$$

The derivatives take the form $\partial \boldsymbol{d}_{\rm tot,i}/\partial \boldsymbol{p}_{s_m\mu} = \partial \boldsymbol{d}_{\rm tot,i}/\partial \boldsymbol{p}_{\rm j,\mu}$ if $m = i$ and are zero otherwise. The matrix $\mathbf{C}_{\rm data}$ denotes the fiducial (KiDS-450) data covariance. Once the Fisher information matrix is calculated for each of the splits used in our analysis, we can calculate its inverse and use it as an estimate of the parameter covariance matrix. This includes now CCs across the parameter sets in equation (B2).

We draw 10^4 samples from a multivariate Gaussian distribution centred on the best-fitting parameters, $\mathbf{p}_{\rm tot,bf}$, and with covariance $\mathbf{F}^{-1}$, cut to within the prior ranges of the original MCMC runs. These parameter samples are then translated into the corresponding theoretical ξ_+ and ξ_- correlation functions and constitute now correlated joint and split TPDs. We use these correlated samples to estimate the correlation coefficients, $r_{\rm ij}$, to propagate the correlations induced by using the same data vector into the final covariance of the difference uncertainties:

$$(\mathbf{C}^{\rm final}_{\rm diff})_{\rm ij} = r_{\rm ij}\sqrt{(\mathbf{C}_{\rm diff})_{\rm ii}\,(\mathbf{C}_{\rm diff})_{\rm jj}}, \tag{B4}$$

where we use the entries from the added covariances of the original 'independent' joint and split MCMC runs:

$$\mathbf{C}_{\rm diff} = \mathbf{C}_{\rm joint} + \mathbf{C}_{\rm split}, \tag{B5}$$

thereby avoiding the simplifications inherent to the Fisher matrix approach for the variances.

The inverse of the matrix $\mathbf{C}^{\rm final}_{\rm diff}$ enters in the calculation of the χ^2 when fitting the differences of the joint and split TPDs to zero on which we base the estimate of the significances for the TPD-based tension estimator (see Section 2.3). Because we are predicting a typically ∼100D distribution from a ∼10D posterior distribution, the estimated covariance matrix $\mathbf{C}^{\rm final}_{\rm diff}$ is necessarily close to being fully correlated. Therefore, we employ a PCA to infer its inverse. We keep only the principal components that contain at least 95 per cent of the total variance to construct the invertible PCA-based covariance, which is then used in the fitting process instead of $\mathbf{C}^{\rm final}_{\rm diff}$.

APPENDIX C: IMPACT OF NOISE AND CORRELATIONS ON GOODNESS-OF-FIT SIGNIFICANCE

In Section 3, we introduced an intuitive criterion to quantify tension between two distributions, such as a TPD and a data distribution. We showed that the criterion only weakly depends on the relative widths of the data and TPD distributions, so that we will fix it in the following to the realistic value of $t/s = 0.1$.

In real data, two sources of noise have to be taken into consideration. First, we usually only have a single realization of the data vector for analysis, which is an unbiased (for Gaussian-distributed data at least) but noisy estimate of the expected value of the data. This leads to random shifts of the data away from the TPD modes, which are indistinguishable from systematic shifts. Consequently, the estimated tension incurs a statistical uncertainty, which is illustrated in Fig. C1, for which we have drawn multiple

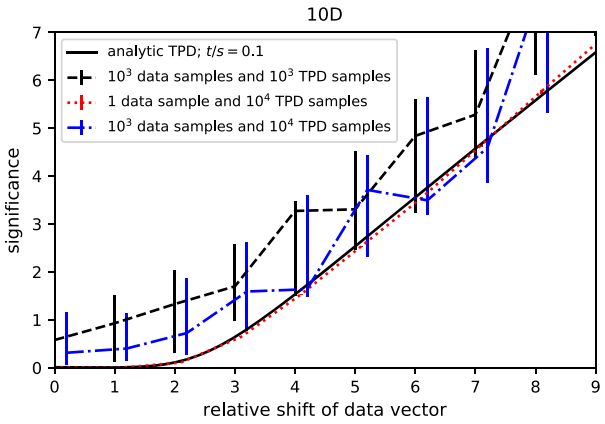

Figure C1. The impact of two sources of noise for the TPD-based tension estimator for a 10D mock data vector: data realization noise and the finite sample size from which the TPDs are constructed. The black dashed line with error bars, for example, is derived for 10^3 random realizations of the data vector and uses also only 10^3 samples for the TPDs. The blue dotted–dashed line with error bars is comparable to the previous one but shows the impact of increasing the number of TPD samples by a factor of 10. Finally, the red dashed line with (tiny) error bars employs only one particular data realization and 10^4 TPD samples. This set-up matches the analytic expectation without these noise effects (black solid line) very well. Note that we apply a small x offset between the black and blue dashed lines to facilitate the comparison of error bar sizes.

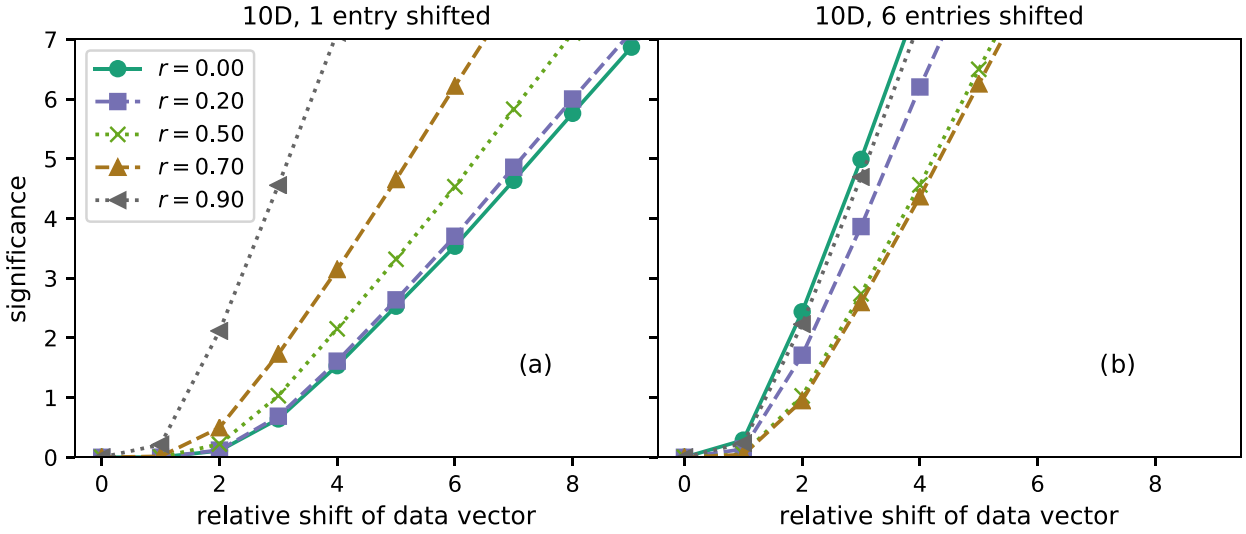

Figure C2. (a) Significance as measured by the TPD tension criterion for a 10D toy model vector for which the first entry was shifted by the amount indicated on the x-axis. Differently coloured lines indicate the chosen value for the correlation parameter r (see equation 52). (b) The same as in (a), but for shifting the first six entries of the data vector at once by the indicated amounts.

realizations of $\boldsymbol{d}_{\rm fid}$ from $\mathcal{N}(\boldsymbol{d}_{\rm fid}; 0, \mathbf{C})$. At low significance, i.e. when data distribution and TPD nearly perfectly overlap, the noise tends to always increase the shifts and thus the tension estimate, which skews the result away from the expectation towards larger values. However, it mostly affects low tension at $\sim 2\,\sigma$ and below, and leads to a conservative conclusion on tension in the data.

The second source of noise is the finite sample size drawn from the TPD. Especially for large tension when only the extreme tails of the data distribution and TPD overlap, the estimate is driven by the single TPD sample that is closest in Euclidean distance to the core of the data distribution. If the TPD sample size is small, the tails are less well covered by sample points, and therefore the tension tends to be overestimated. This trend can also be seen in Fig. C1 for a change from 10^3 to 10^4 samples; however, the difference is small, so that we consider 10^4 TPD samples for the real-data case as sufficiently stable. Note that, modulo these noise effects, the TPD measurement pipeline reproduces the analytic expectation well.

The correlation functions used in the KiDS-450 data analysis feature strong CCs between angular bins, tomographic redshift bin combinations, as well as between ξ_+ and ξ_-. This impedes any attempt at spotting discrepancies with best-fitting models or the TPDs 'by eye'. To demonstrate the effect of correlations, we introduce $r > 0$ into our toy model for both the data distribution and TPD, with results shown in Fig. C2. If one data point is perturbed, positive correlations increase the tension, as large values of r imply that the data strongly prefer that data points lie on the same side of the model (in our case, the model is zero everywhere). As more and more data points are shifted (by equal amounts q in our toy model), the largest values of r lead to a decrease in tension because the systematic shift lines up the data points as preferred by the correlation structure of the data covariance.

APPENDIX D: ADDITIONAL FIGURES FOR THE KIDS-450 ANALYSIS

In Section 5.2, we presented the ξ_+ correlation functions for all unique tomographic bin combinations from KiDS-450 in combination with the means of the joint and split TPDs for all four splits into subsets. This serves for visual comparison of data to TPDs and also to guide the reader's intuition for the concept of TPDs. For completeness, we show here in Fig. D1, the corresponding figures for the ξ_- correlation functions.

We also investigate the significances derived with both TPD-based estimators for all four splits of the KiDS-450 data per ξ_+ (left-hand panels) and ξ_- (right-hand panels) correlation function for all tomographic bin combinations. Those are presented in Figs D2 and D3, respectively. In both figures, we also present in the columns labelled with 'no B' the results when subtracting off the measured small-scale B modes in KiDS-450. The different colours correspond to one particular tomographic bin combination $z_{\rm i} \times z_{\rm j}$ as indicated in the legends of the left-hand panels. In Fig. D2, the open circles correspond to the significance estimates derived from the joint TPDs, whereas the crosses are derived from the split TPDs.

The features to highlight in Figs D2(a) and (b) are the constant levels of significance of $\sim 2.5\,\sigma$ and $\sim 1.8\,\sigma$ for the $z_2 \times z_4$ and $z_1 \times z_2$ tomographic bin combinations (almost) independent of the particular split into subsets. This is a strong hint for that the mismatch between data and theory in these two tomographic bins is driving the bad goodness of fit reported already in Fig. 8(a). Parts of that mismatch seem to be driven by a residual systematic between small and large angular scales and the small-scale B modes in the ξ_+ correlation function for $z_2 \times z_4$ as the significances derived from the split TPD (crosses) are lower than the ones based on the joint TPD. The significances shown in Figs D3(a) and (b) are the highest in ξ_+ ($\lesssim 1.5\,\sigma$) for the split into large versus small angular scales (i.e. 'LS versus SS'). However, the $z_2 \times z_4$ tomographic bin combination does not stand out in this estimator and instead the largest contribution comes from the $z_4 \times z_4$ combination.

Subtracting off the small-scale B modes decreases the tension in every tomographic bin combination to $\lesssim 0.5\,\sigma$, which is also consistent with the results for the other three splits. As mentioned in Section 5.2, it is also interesting to point out that the significances of all splits decrease when subtracting off the small-scale B modes, except for the split into 'z-bin 4 (and all its CCs) versus all' (other z-bin combinations). The increase in significance is small for each individual correlation function but occurring simultaneously for all. Therefore, we interpret this behaviour as a sign for that removing the small-scale B modes from the fiducial data vector pronounces a residual systematic in the 'z-bin 4 versus all' split.

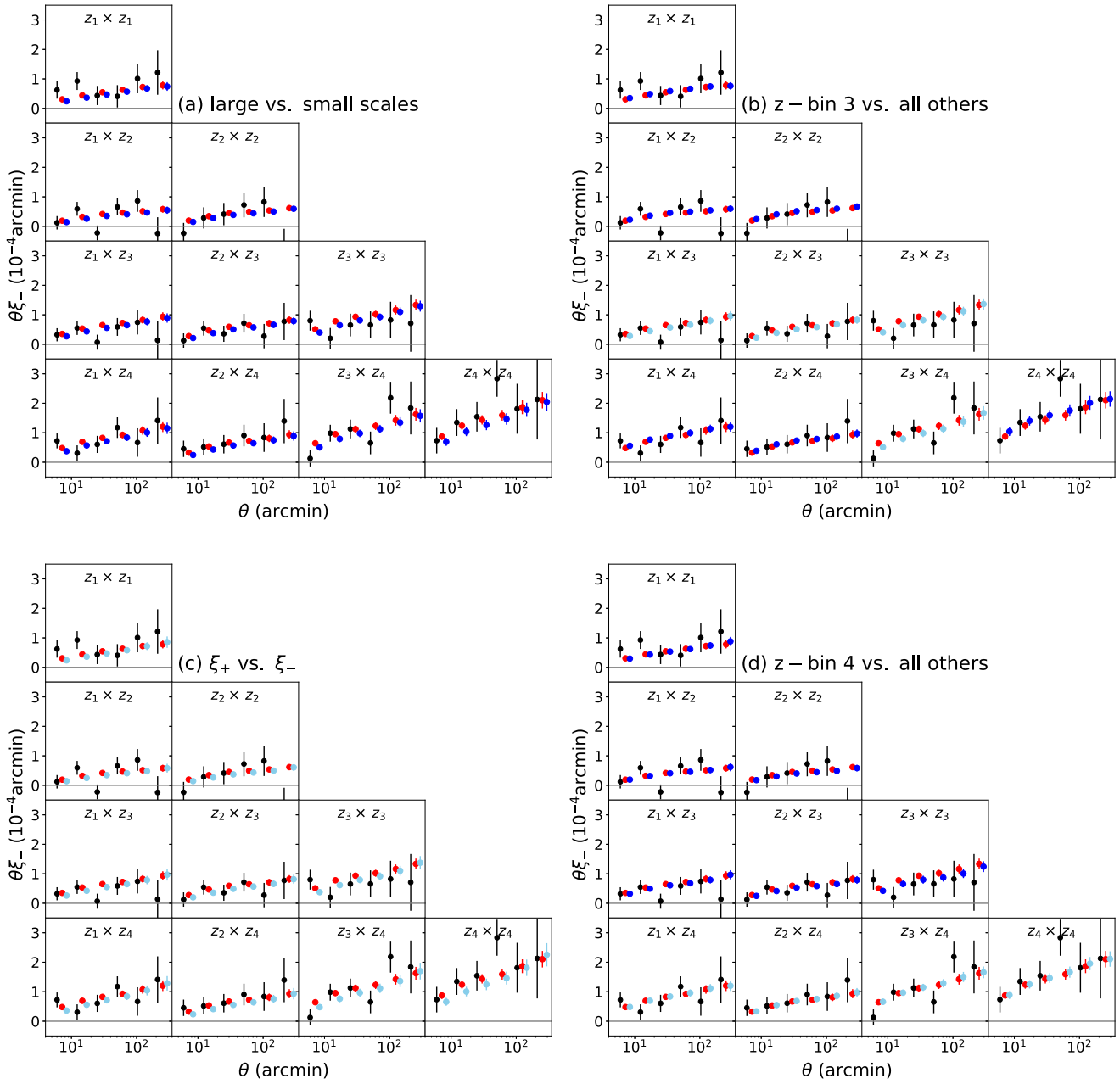

Figure D1. KiDS-450 data vector (black points) and TPD means (red and blue points) for the ξ_- estimator as a function of angular scale per redshift bin combination. The TPDs are based on the joint (red points; the same in all panels) and split (dark and light blue points) cosmological and nuisance parameters from MCMC runs fit to the data vector. The panels consider various mutually exclusive splits of the data vector (a) large versus small scales, (b) z-bin 3 (and all its CCs) versus all other redshift correlations, (c) ξ_+ versus ξ_-, and (d) z-bin 4 (and all its CCs) versus all other redshift correlations. Error bars on the means are derived from the 68 per cent credibility interval around the mean. The erro rbars for the data are based on the diagonal of the covariance matrix.

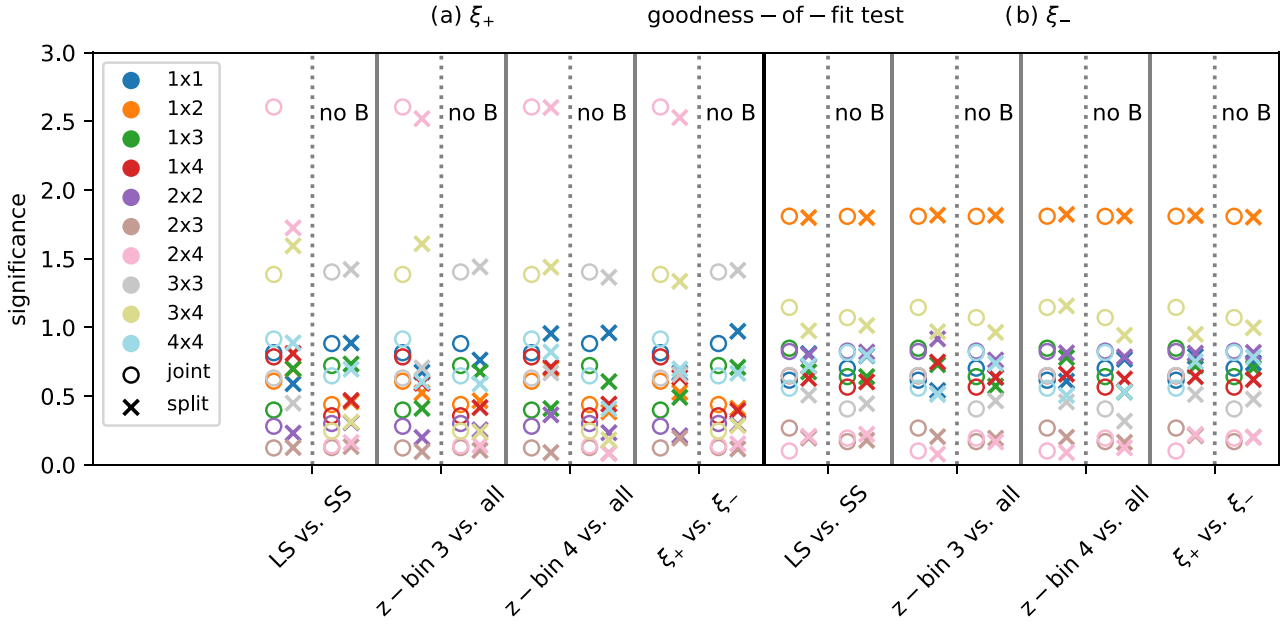

Figure D2. Significances for the goodness of fit estimated by comparing the joint (open circles) and split TPDs (crosses) to the data distribution for the following splits of the fiducial KiDS-450 data vector (from left to right): large versus small scales ('LS versus SS'), z-bin 3 and all its CCs versus all other z-bin combinations ('z-bin 3 versus all'), z-bin 4 and all its CCs versus all other z-bin combinations ('z-bin 4 versus all'), 'ξ_+ versus ξ_-'. (a) Using the ξ_+ estimator only and per tomographic bin combination $i \times j$. (b) Using the ξ_- estimator only and per tomographic bin combination $i \times j$. The columns marked with 'no B' use the KiDS-450 data vector from which the measured small-scale B modes were removed.

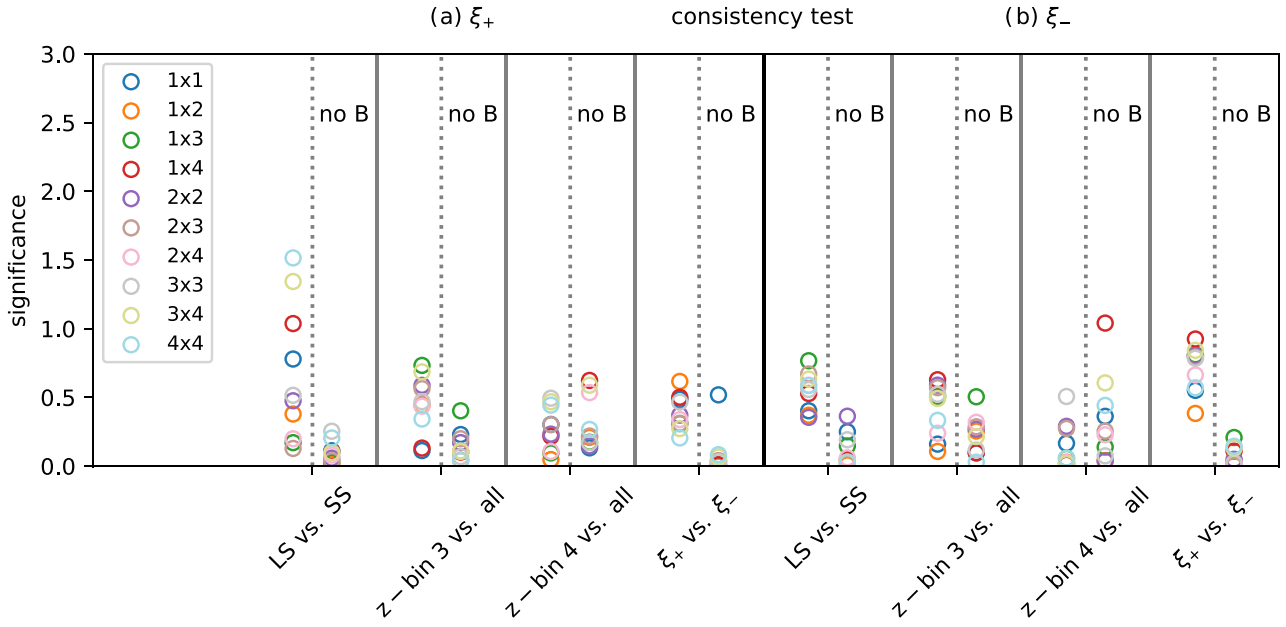

Figure D3. Significances for tension estimated by comparing the differences of the joint and split TPDs to zero for the following splits of the fiducial KiDS-450 data vector (from left to right): large versus small scales ('LS versus SS'), z-bin 3 and all its CCs versus all other z-bin combinations ('z-bin 3 versus all'), z-bin 4 and all its CCs versus all other z-bin combinations ('z-bin 4 versus all'), 'ξ_+ versus ξ_-'. (a) Using the ξ_+ estimator only and per tomographic bin combination $i \times j$. (b) Using the ξ_- estimator only and per tomographic bin combination $i \times j$. The columns marked with 'no B' use the KiDS-450 data vector from which the measured small-scale B modes were removed.

This paper has been typeset from a TEX/LATEX file prepared by the author.