Quantifying Suspiciousness within correlated data sets

Pablo Lemos [1]★ Fabian Köhlinger,[2] Will Handley [3,4] Benjamin Joachimi,[1] Lorne Whiteway[1] and Ofer Lahav[1]

[1]*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2]*Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan*
[3]*Astrophysics Group, Cavendish Laboratory, J. J. Thomson Avenue, Cambridge CB3 0HE, UK*
[4]*Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK*

ABSTRACT

We propose a principled Bayesian method for quantifying tension between correlated data sets with wide uninformative parameter priors. This is achieved by extending the Suspiciousness statistic, which is insensitive to priors. Our method uses global summary statistics, and as such it can be used as a diagnostic for internal consistency. We show how our approach can be combined with methods that use parameter space and data space to identify the existing internal discrepancies. As an example, we use it to test the internal consistency of the KiDS-450 data in four photometric redshift bins, and to recover controlled internal discrepancies in simulated KiDS data. We propose this as a diagnostic of internal consistency for present and future cosmological surveys, and as a tension metric for data sets that have non-negligible correlation, such as Large Synoptic Spectroscopic Survey and *Euclid*.

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

1 INTRODUCTION

Quantifying consistency between different data sets has become one of the main challenges in modern cosmology. With the increasing number of methods and surveys measuring different properties of the Universe, it is crucial to develop appropriate statistical tools to compare and combine these data sets. This is important for two reasons. First, differences in cosmological constraints from different data sets, i.e. 'tensions', could indicate unaccounted-for systematic errors in one or both data sets, or could indicate that the underlying theoretical model (for example lambda cold dark matter, ΛCDM) is not sufficient to explain both data sets. Secondly, because the combination of different data sets can provide powerful cosmological constraints by breaking degeneracies existing in single data sets, but data sets can only be combined meaningfully if they are in agreement.

Cosmological tensions and their correct quantification are now particularly important, given that while all existing data sets match the ΛCDM model of cosmology individually, there are two main disagreements between data sets that could be a hint of beyond ΛCDM physics. This belief is reinforced by the fact the tensions are between high-redshift measurements of the cosmic microwave background by the *Planck* satellite (Planck Collaboration XIII 2015, 2018), and low-redshift measurements of the Hubble constant (Riess et al. 2018, 2019) and of the growth of structure measured by galaxy clustering and weak gravitational lensing by several surveys such as the Canada–France–Hawaii Telescope Lensing Survey (Heymans et al. 2012, 2013; Joudaki et al. 2017), the Kilo-Degree Survey

(KiDS; Hildebrandt et al. 2017; Köhlinger et al. 2017), and the Dark Energy Survey (DES; Abbott et al. 2018; Troxel et al. 2018a).

There is a recent and extensive literature on quantifying tension in the context of cosmology (for a review of some methods, see Charnock, Battye & Moss 2017). Perhaps the most frequently used is the Bayes ratio R introduced in Marshall, Rajguru & Slosar (2006), which has the advantages of using exclusively Bayesian quantities and of being parametrization independent. However, this method has the disadvantage of being proportional to the prior volume (which can hide existing tensions when broad priors are chosen with the goal of being uninformative). Other approaches based on differences in the best-fitting parameters were introduced by Lin & Ishak (2017a, b), Raveri & Hu (2019), and Adhikari & Huterer (2019). Kunz, Trotta & Parkinson (2006), Seehars et al. (2014), Grandis et al. (2016a, b), and Nicola, Amara & Refregier (2019) used methods based on the Bayesian information measured by the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951). One of the latest methods suggested was the Bayesian 'Suspiciousness' introduced in Handley & Lemos (2019b, henceforth H19). The Suspiciousness acts as an extension of the R statistic of Marshall et al. (2006) to the case of uninformative priors, preserving many of the desired properties of R without having dependence on prior volume. It is important to point out that both R and S rely on quantities, such as the Bayesian evidence, that are noisy statistics. This was already discussed by Jenkins & Peacock (2011), and will be further investigated in Joachimi (2019, in preparation).

However, all the methods above assume that the data sets are uncorrelated. In this work, we focus on methods to quantify consistency between correlated data sets [i.e. data sets for which $P(A, B) \neq P(A)P(B)$]. The measurement of tension in correlated data sets
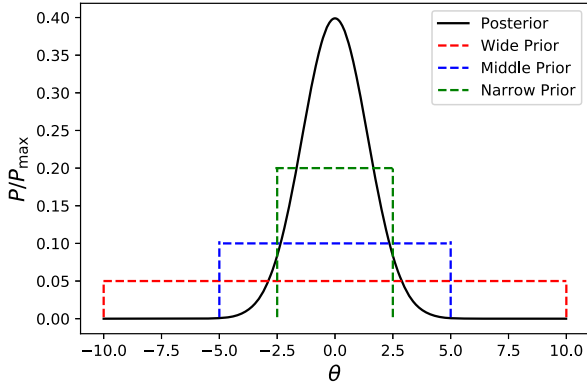
★ E-mail: pablo.lemos.18@ucl.ac.uk

Figure 1. An example of different priors for a Gaussian distribution. The 'wide prior' in red contains all the posterior volume but extends far into the tails. The 'middle prior' in blue, is an approximation that also contains the posterior volume but not the tails. The 'narrow prior' significantly changes the shape of the posterior distribution, as it 'chops' it where it is non-negligible. The Suspiciousness statistic would not differentiate between 'middle' and 'wide' priors, unlike the Bayes ratio. However, both statistics would obtain different results for the 'narrow' prior, which is the desired behaviour as this prior alters the shape of the posterior.

has numerous applications: it can be used to quantify consistency between different 'splits' or parts of a given data set, as a measure of internal consistency. It is also relevant when combining different data sets with non-trivial correlations. For example, this will be necessary to obtain combined constraints from *Euclid* (Laureijs et al. 2011) and the Large Synoptic Spectroscopic Survey (LSST; Ivezić et al. 2019), which as showcased by Rhodes et al. (2017) and Schuhmann, Heymans & Zuntz (2019) would be very well motivated.

The problem of assessing consistency between correlated data sets was tackled by Köhlinger et al. (2019, henceforth K19), in the context of the KiDS-450 data. They use three 'tiers' of consistency tests: Bayesian evidence ratios, parameter differences, and posterior predictive distributions (PPD). They conclude that Bayesian evidence ratios are not appropriate for this problem because of their dependence on prior volume. As already mentioned, there are several important advantages to the use of Bayesian evidence ratios with respect to the other two tiers of consistency test introduced in K19. Therefore, in this paper, we extend the work of H19 and apply the Suspiciousness statistic to the case of correlated data sets with broad non-informative priors. This provides a Bayesian, parametrization-independent measure of consistency between correlated data sets.

In Section 2, we describe how the Suspiciousness can be extended to correlated data sets. Section 3 applies this method to a toy model consisting of Gaussian data sets that we 'split' in two. We discuss how this method is related to the other two tiers of internal consistency introduced in K19 in Section 4. The method is applied to the KiDS-450 data in Section 5. We present our conclusions in Section 6.

2 TENSION IN CORRELATED DATA SETS

The goal of this section is to extend the tension metrics and Bayesian model dimensionality (BMD) introduced in Handley & Lemos (2019a, b) to the case of quantifying tension in parameters from posteriors that come from correlated data sets. To derive the R statistic, Marshall et al. (2006) propose comparing two hypotheses:

(i) H_0: There exists one common set of parameters that describes both data sets.
(ii) H_1: There exist two sets of parameters, one for each data set.

Bayes' theorem is then used to calculate the ratio of degrees of belief in each hypothesis in light of the data D:

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(D|H_0)}{P(D|H_1)} \cdot \frac{P(H_0)}{P(H_1)} \propto \frac{\mathcal{Z}_0}{\mathcal{Z}_1} = R, \quad (1)$$

where R is the Bayes ratio and $\mathcal{Z}$ is the Bayesian evidence

$$\mathcal{Z}_i \equiv P(D|H_i) = \int d\theta \mathcal{L}(\theta)\Pi(\theta), \quad (2)$$

with θ the parameters of the model, $\mathcal{L}(\theta) \equiv P(D|\theta, H)$ the likelihood, and $\Pi(\theta) \equiv P(\theta|H)$ the prior. If our prior belief in both hypotheses is the same $P(H_0) = P(H_1)$, the proportionality constant in equation (1) becomes unity. Under this formulation, we may view tension quantification as a model comparison problem with R as a figure of merit. Finally, the posterior from Bayes theorem and the KL divergence (Kullback & Leibler 1951) are defined via

$$P(\theta|D, H) \equiv \mathcal{P}(\theta) = \frac{\mathcal{L}(\theta)\Pi(\theta)}{\mathcal{Z}}, \quad (3)$$

$$\mathcal{D} = \int d\theta \, \mathcal{P} \log\left(\frac{\mathcal{P}}{\Pi}\right). \quad (4)$$

The Bayes ratio approach using R to quantify tension for correlated data sets was used in K19, but, as they show, R depends on the prior volume. This is the same prior volume dependence that is extensively discussed in H19 for uncorrelated data sets. While this would not be a problem for well-motivated priors, it means that we cannot rely on Bayesian evidence ratios obtained with priors that are purposefully chosen to be broad (with the intention of being uninformative).

To address these concerns about prior volume, we can instead use the Suspiciousness S introduced in H19, which can be understood as the value of R that corresponds to the narrowest possible priors that do not significantly alter the shape of the posteriors, i.e. that contain nearly the entire volume of the posterior distribution, which would approximately correspond to the 'middle prior' in Fig. 1. The natural logarithm of the Suspiciousness is given by

$$\log S = \log R - \log I. \quad (5)$$

The information I quantifies the a priori probability that the data sets would match given the prior range. The larger the prior range relative to the posterior constraints, the lower the probability that the constraints will be consistent. For correlated data sets, the natural logarithm of the information is given by

$$\log I = \mathcal{D}_1 - \mathcal{D}_0. \quad (6)$$

In the uncorrelated case, the additivity of the KL divergence implies that $\mathcal{D}_1 = \mathcal{D}(A) + \mathcal{D}(B)$ and $\mathcal{D}_0 = \mathcal{D}(A, B)$, which recovers the methodology of H19.

3 GAUSSIAN EXAMPLE

To illustrate the formalism, we consider the example of multivariate Gaussian posterior distributions (Fig. 2). Let A and B be two data sets that can constrain the same set θ of d parameters. Each data set A or B individually constrains one set of parameters θ_A or θ_B giving a posterior with parameter mean and covariance of (μ_A, Σ_A) or (μ_B, Σ_B), respectively.

Assume now that the two data sets are correlated. If we combine the two data sets and use hypothesis H_1, then the mean and covariance of the marginal distribution of each individual set of parameters will be unchanged; however, the data correlation will induce a correlation
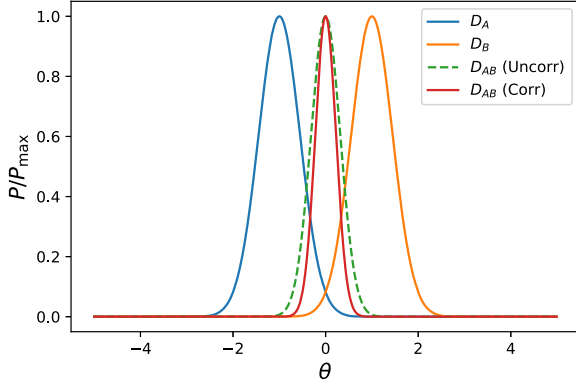
Figure 2. Graphical illustration of the Gaussian example, in the case of a single parameter. The plot shows the two posteriors, and their combination in the uncorrelated and correlated case, for $\sigma_A^2 = \sigma_B^2 = 0.1$, and $V_\pi = 10$.

between parameter sets so that the likelihood will be

$$\log \mathcal{L}_1 = \log \mathcal{L}_1^{\max} - \frac{1}{2} \begin{bmatrix} \theta_A - \mu_A \\ \theta_B - \mu_B \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Sigma_A & \Sigma_X \\ \Sigma_X^\mathsf{T} & \Sigma_B \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} \theta_A - \mu_A \\ \theta_B - \mu_B \end{bmatrix}, \qquad (7)$$

where the cross matrix Σ_X will be zero in the uncorrelated case.

The likelihood for hypothesis H_0 can be found by setting $\theta_A = \theta_B = \theta$ in equation (7). Notationally, it is helpful to define the precision matrix Λ in block form via

$$\Sigma_1 = \begin{bmatrix} \Sigma_A & \Sigma_X \\ \Sigma_X^\mathsf{T} & \Sigma_B \end{bmatrix}, \qquad \Sigma_1^{-1} = \Lambda = \begin{bmatrix} \Lambda_A & \Lambda_X \\ \Lambda_X^\mathsf{T} & \Lambda_B \end{bmatrix}, \qquad (8)$$

where the different blocks in Λ are related to those in Σ_1 by Schur's complement (Woodbury 1950; Zhang 2005). After some effort manipulating matrix expressions (see Appendix), we find

$$\log \mathcal{L}_0 = \log \mathcal{L}_0^{\max} - \frac{1}{2} (\theta - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\theta - \mu_0), \qquad (9)$$

$$\log \mathcal{L}_0^{\max} = \log \mathcal{L}_1^{\max} - \frac{1}{2} (\mu_A - \mu_B)^\mathsf{T} \Sigma_{\Delta\mu}^{-1} (\mu_A - \mu_B), \qquad (10)$$

$$\Sigma_{\Delta\mu} = \Sigma_A - \Sigma_X - \Sigma_X^\mathsf{T} + \Sigma_B, \qquad (11)$$

$$\mu_0 = \Sigma_0 \left[\left(\Lambda_A + \Lambda_X^\mathsf{T} \right) \mu_A + (\Lambda_X + \Lambda_B) \mu_B \right], \qquad (12)$$

$$\Sigma_0 = \left(\Lambda_A + \Lambda_X + \Lambda_X^\mathsf{T} + \Lambda_B \right)^{-1}. \qquad (13)$$

As discussed in H19, the evidence and KL divergence for a multivariate Gaussian with a flat prior of volume V_Π that essentially completely encloses the posterior (Fig. 1) are

$$\log \mathcal{Z} = \log \mathcal{L}^{\max} + \frac{1}{2} \log |2\pi \Sigma| - \log V_\Pi, \qquad (14)$$

$$\mathcal{D} = \log V_\Pi - \frac{1}{2} (d + \log |2\pi \Sigma|). \qquad (15)$$

Combining equations (1),(6),(5),(7),(9),(10),(15), and (14) yield

$$\log R = \log V_\Pi + \log \mathcal{L}_0^{\max} - \log \mathcal{L}_1^{\max} + \frac{1}{2} \log \frac{|2\pi \Sigma_0|}{|2\pi \Sigma_1|}, \qquad (16)$$

$$\log I = \log V_\Pi - \frac{d}{2} + \frac{1}{2} \log \frac{|2\pi \Sigma_0|}{|2\pi \Sigma_1|}, \qquad (17)$$

$$\log S = \frac{d}{2} - \frac{1}{2} (\mu_A - \mu_B)^\mathsf{T} \Sigma_{\Delta\mu}^{-1} (\mu_A - \mu_B). \qquad (18)$$

Note that a factor V_Π emerges in R since H_1 has twice as many parameters as H_0, but that this dependence on prior volume is mirrored in I which therefore cancels in $\log S$. It can easily be verified that if the data are uncorrelated ($\Sigma_X = \Lambda_X = 0$) then these results agree with those of H19.

Now under hypothesis H_0, $\mu_A - \mu_B$ will be distributed with mean zero and covariance $\Sigma_{\Delta\mu}$, and thus $(\mu_A - \mu_B)^\mathsf{T} \Sigma_{\Delta\mu}^{-1} (\mu_A - \mu_B)$ will have a χ_d^2 distribution (as can be shown by Cholesky decomposing the covariance matrix). Thus $d - 2\log S$ must follow a χ_d^2 distribution as well (as was the case in H19). The probability p_t of the data sets being discordant by chance is

$$p_t = \int_{d-2\log S}^{\infty} \chi_d^2(x) \, \mathrm{d}x = \int_{d-2\log S}^{\infty} \frac{x^{d/2-1} \mathrm{e}^{-x/2}}{2^{d/2} \Gamma(d/2)} \, \mathrm{d}x. \qquad (19)$$

The effective number of dimensions constrained by both data sets d is given by the BMD introduced in Handley & Lemos (2019a). In particular, Handley & Lemos (2019a) show that d is exactly the same as the number of dimensions for a Gaussian likelihood. For correlated data sets, d is given by

$$d = d_1 - d_0, \qquad (20)$$

where the number of parameters constrained simultaneously by both data sets is given by subtracting the number of parameters constrained by the combination of both from the number of parameters constrained by each data set separately, in a similar manner to Handley & Lemos (2019a) for uncorrelated data sets.

As demonstrated by H19, while the results of this section have been obtained for Gaussian likelihoods, they can be applied to more general posteriors. This is because the Suspiciousness is invariant under coordinate transformations, such as Box–Cox transformations (Box & Cox 1964; Joachimi & Taylor 2011; Schuhmann, Joachimi & Peiris 2016), which 'Gaussianize' the posterior. Furthermore, H19 showed that the Suspiciousness recovers the intuitively correct answers for cosmological examples, which are somewhat non-Gaussian, particularly in the nuisance parameters.

4 CONNECTION TO OTHER INTERNAL CONSISTENCY TESTS

K19 proposed three tiers of internal consistency between data sets. We have previously discussed how the Bayesian Suspiciousness serves as an alternative to tier 1 (the Bayes ratio). In this section, we discuss the connection between Suspiciousness and tiers 2 and 3. We argue that, while these three methods are different in their implementation, the underlying quantities being calculated are surprisingly similar.

Tier 2 of K19 are differences of parameter duplicates: after calculating the posterior distributions under hypothesis H_1 (the data sets are described by different sets of parameters), we can derive the posterior distribution for the differences between the parameters describing each data set. K19 then proposes finding the fraction of samples with a value of the posterior smaller than the value in the origin (which is the point of perfect agreement). This can be seen as an extension of tension metrics based on parameter shifts, such as those introduced in Raveri & Hu (2019) and Adhikari & Huterer (2019), to the case of correlated data sets. From the result of Section 3, it is easy to see how our method is connected to this. For Gaussian posterior distributions, we show that $d - 2\log S$ is χ^2 distributed, from which we can calculate the tension probability. While the methodology is different, the quantity being calculated is the same in the case of this Gaussian toy model. In fact, one
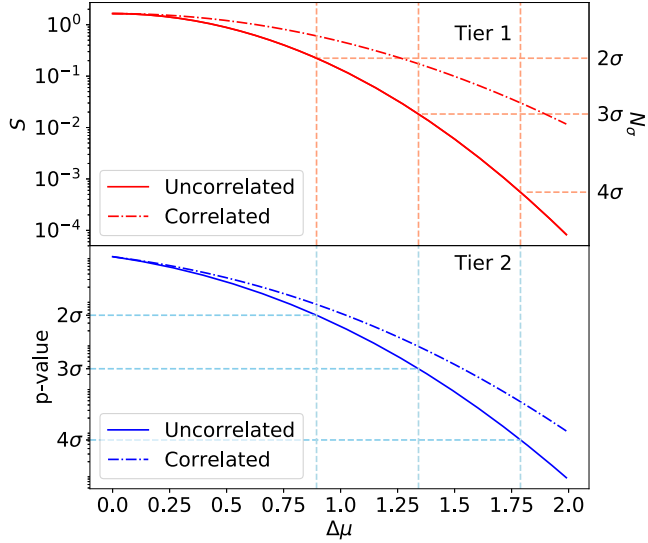
Figure 3. Comparison of tiers 1 and 2 of statistical consistency for the toy model of Fig. 2. The top panel shows the Suspiciousness, and the bottom panel shows the p value obtained from tier 2 of K19 (parameter differences). The *x*-axis shows $\Delta\mu \equiv |\mu_A - \mu_B|$, where μ_A and μ_B are the means of the Gaussian distributions in the toy model in Section 3. The solid curves correspond to uncorrelated Gaussians, while the dash–dotted curve corresponds to the correlated case. The dashed horizontal and vertical lines show the 2σ, 3σ, and 4σ values for the uncorrelated cases, which agree exactly in this case.
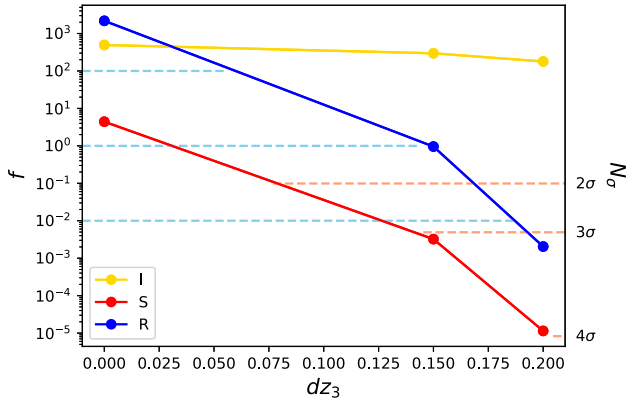

Figure 4. The evidence ratio (R, in blue) and the two factors into which it can be split, according to H19: the information (I, in yellow) and the Suspiciousness (S, in red). The *y*-axis on the left is the value of each function, which in the case of R corresponds to a ratio of probabilities. The *y*-axis on the right is the number of sigma corresponding to different values of S (note that the positions of the lines depend on the effective number of constrained directions). The *x*-axis is different values of dz_3 (i.e. shifts in the source redshift distribution of bin 3) in the sensitivity analysis. The blue line is similar to the left-hand panel in fig. A1 of K19.

could argue that the '$m\sigma$' interpretation presented in K19 should be dependent on dimensionality, which could be calculated using the BMD (Handley & Lemos 2019a), and then the calculation of '$m\sigma$' would become similar to equation (19). This is shown in Fig. 3, where we compare the results of our method and parameter differences for different 1D Gaussian distributions, and confirm that the results are the same.

Tier 3 of K19 consists of using the PPD. This technique was introduced in Gelman, li Meng & Stern (1996), and has been applied to cosmology in recent problems (e.g. Abbott et al. 2019; Feeney

et al. 2019). When using the PPD, we calculate the probability of data D_A conditional on data D_B and the underlying model M in the following way:

$$P(D_A|D_B, M) = \int d\theta \; P(D_A|\theta, M) P(\theta|D_B, M). \quad (21)$$

In practice, this means that from samples of the posterior for D_B, we calculate realizations of the likelihood for D_A. The most challenging part of using the PPD is calibrating the distribution. $P(D_A|D_B, M)$ is a probability density that is challenging to normalize. The authors of K19, for example, built a 'Translated Probability Distribution' from a predicted data vector, with the goal of calibrating the PPD. However, there is an alternative way of interpreting this number, which is by taking the ratio $P(D_A|D_B, M)/P(D_A|M)$. This quantity is unitless, and can therefore be interpreted as a probability ratio. More importantly, as discussed in Handley & Lemos (2019b), this ratio is equal to the Bayes ratio R. In this way, we can see tiers 1 and 3 of K19 as being elements of the same calculation, but using different calibrations of the PPD. If both methods are correct, then they should produce similar results.

5 KIDS-450 DATA

In this section, we test the methods introduced in Section 2 on the KiDS-450 data (Hildebrandt et al. 2017).[1] KiDS uses the correlations in the shapes of galaxy images to measure weak gravitational lensing caused by large-scale structure (for reviews, see Bartelmann & Schneider 2001; Kilbinger 2015). Their measurements span a redshift range $z = [0.1, 0.9]$; this range is divided into four redshift bins each of width $\Delta z = 0.2$. The estimators used are the correlation functions $\xi_+(i, j)$, $\xi_-(i, j)$, with seven and six angular bins, respectively, and where $i, j = 1, .., 4$ refer to the redshift bins. The data therefore consist of 130 data points [note that $\xi_\pm(i, j) = \xi_\pm(j, i)$].

We have chosen to examine the KiDS data because Efstathiou & Lemos (2018, henceforth E18) found possible inconsistencies between different splits of the data using two simple statistical tests. In particular, bins 3 and 4 were found to be inconsistent with the rest of the data with significances 2.60σ and 3.52σ. Following this work, K19 performed their three-tier consistency tests on the same data. Using these statistics, K19 found that the significance of these inconsistencies is reduced, and concluded that the results of E18 are sensitive to the overall goodness of fit of the data. Following E18, Troxel et al. (2018b) improved the shot noise model of the analytically derived covariance matrix by including a model for the survey-boundary effects. This correction increased the size of the error bars at large angular scales, reducing the discrepancies found in E18 to below significant levels. Asgari et al. (2019) also quantified the errors associated with each of the definitions of the angular tomographic bins centre used in Troxel et al. (2018b) and Hildebrandt et al. (2017), proposing an accurate new approach adopted in Hildebrandt et al. (2020). In the remainder of this work, however, we will use the data from Hildebrandt et al. (2017) as analysed by both E18 and K19. These corrections increased the size of the error bars at large angular scales, reducing the discrepancies found in E18 to below significant levels. However, in the remainder of this work, we will use the data from Hildebrandt et al. (2017) used by both E18 and K19.

---

[1]See http://kids.strw. leidenuniv.nl/sciencedata.php.

Table 1. Comparison of evidence $\log R$ (interpreted with the Jeffreys scale used in K19) and Suspiciousness $\log S$ for some redshift shifts in the sensitivity analysis. The last three columns show the BMD d, the tension probability p_t and the corresponding number of sigma, calculated as $N_\sigma \equiv \sqrt{2}\mathrm{Erf}^{-1}(1 - p_t)$.

dz_3	$\log R$	Interpretation	$\log S$	d	p_t	N_σ
0	7.69 ± 0.15	Decisive Agreement	1.488 ± 0.049	2.68 ± 0.17	0.99995 ± 0.00053	0
0.15	-0.04 ± 0.14	Neutral	-5.735 ± 0.052	2.75 ± 0.18	0.00203 ± 0.00028	3.089 ± 0.041
0.20	-6.18 ± 0.14	Decisive Tension	-11.383 ± 0.051	2.60 ± 0.17	$7.6 \cdot 10^{-6} \pm 1.4 \cdot 10^{-6}$	4.478 ± 0.039

For our analysis, we compare the KiDS data vectors that involve redshift bin 3 to all others, since bin 3 was one of the discrepant bins in E18. We use the code 2cosmos[2] to sample the KiDS likelihood. 2cosmos is an extension of the public code montepython[3] (Audren et al. 2013). We compute evidence ratios, information, and Suspiciousness using the public code anesthetic[4] (Handley 2019). We first reproduce the results from table 2 of K19[5] ($\log R = 4.21 \pm 0.15$), and then apply our method to the same split of the data. We get a log Suspiciousness $\log S = -1.992 \pm 0.064$. Under the Gaussian approximation, we can assign a *tension probability* p_t to this value: we get the number of constrained dimensions $n_d = 2.75 \pm 0.18$, and the corresponding tension probability of the data sets being consistent is $p = 0.0674 \pm 0.0062$, corresponding to a significance 1.83σ.

As expected, in this case the Suspiciousness provides a far more accurate assessment of tension than does R, which obtains 'very strong' evidence for agreement according to the Jeffreys scale used in K19. However, the significance is also lower than in E18, which is a reflection on the dependence on the overall goodness of fit of the statistic used in that paper. In other words, our method quantifies the internal consistency of the data, but is 'blind' to the goodness of fit, while the method used in E18 was sensitive to both. This is a very desirable quality in an internal consistency test, as it allows us to detect the origin of a potential problem in the data. For the KiDS-450 data, the Suspiciousness says that bin 3 is on the edge between consistent and moderately inconsistent, because it measures only the degree of agreement between the predictions of this bin and the rest of the data. However, we stress the need to combine a Suspiciousness internal consistency test with a reliable test of the overall goodness of fit: the Suspiciousness alone would have been insufficient to identify the suboptimal KiDS-450 covariance matrix, while the E18 tests did so precisely because of their sensitivity to the overall goodness of fit.

As a final test of our statistical method, we repeat the sensitivity analysis of appendix A1 in K19: instead of using the real KiDS-450 data, we use mock data vectors, in which the source redshift distribution is shifted by $dz_3 = 0, 0.15, 0.20$. Our results are shown in Fig. 4 and Table 1. Again, we see how the Bayes ratio R can hide tensions because of the width of the priors, while the Suspiciousness shows clear tension for both shifts of the redshift distribution. For example, $dz_3 = 0.15$ corresponds to 'No Evidence' in R, while for the same shift S yields a more than 3σ discrepancy, indicating a strong tension. This highlights the value of the Suspiciousness for internal consistency tests in large data sets. The table also shows how the number of sigma is zero for $dz_3 = 0$. This is because in this case we are comparing data sets whose posteriors overlap perfectly. We should generally also be suspicious of cases where the tension

probability is very close to one (such as this one), as they indicate that the agreement is 'too good'.

6 CONCLUSIONS

Correlations between data sets affect tension measurements, and require specific tension metrics to take into account the full error model, as failing to do so will lead to wrong statements about consistency between data sets. In this paper, we extended the novel tension metric introduced in Handley & Lemos (2019b) to the case of quantifying tension in parameters from posteriors that derive from correlated data sets. This takes the role of 'tier 1' consistency tests introduced in Köhlinger et al. (2019), and provides a measure of consistency for correlated data sets that uses Bayesian quantities, is parametrization independent, and can be used for the case of wide, uninformative priors since it does not depend on the prior volume (in contrast to the Bayes ratio for data set comparison R introduced in Marshall et al. 2006).

We applied this formalism to a Gaussian toy model, and used this to compare it with tiers 2 and 3 of K19, showing that we obtain the same results for Gaussian posteriors. We propose our method as a diagnostic tool of consistency between correlated data sets, that can be complemented by best-fitting calculations (tier 2), and PPD (tier 3) to identify the origin of existing internal tension.

We applied this formalism to the case of KiDS-450 data from Hildebrandt et al. (2017), focusing on tests of tomographic redshift bin 3 versus the rest of the data (bins 1, 2, and 4). We find that the 2.60σ tension detected in Efstathiou & Lemos (2018) reduces to 1.83σ. We interpret this as a difference in the tension metric used: the methods used in Efstathiou & Lemos (2018) depend on the overall goodness of fit, while those used in this work do not. This is a desirable quality, as it allows us to identify the origin of potential problems in the data, but it stresses the need to combine the Suspiciousness with a goodness of fit test.

This method can generally be used for correlated data sets where uninformative priors are used. In particular, it can be used for internal consistency tests of cosmological surveys (such as DES and KiDS), and also to quantify tension between data sets with a non-negligible covariance (such as *Euclid* and LSST). We conclude that due to its Bayesian nature and intuitive interpretation, this method serves as a perfect diagnostic of internal consistency, and we encourage present and future cosmological surveys to use it.

[2] https://github.com/fkoehlin/montepython_2cosmos_public
[3] https://github.com/baudren/montepython_public
[4] https://github.com/williamjameshandley/anesthetic
[5] We use natural logarithms, while K19 use base 10 logarithms.

DATA AVAILABILITY

The data underlying this article are available in the KiDS website, at http://kids.strw. leidenuniv.nl/sciencedata.php.

REFERENCES

Abbott T. M. C. et al., 2018, Phys. Rev. D, 98, 043526
Abbott T. M. C. et al., 2019, Phys. Rev. D, 100, 023541
Adhikari S., Huterer D., 2019, J. Cosmol. Astropart. Phys., 2019, 036
Asgari M. et al., 2019, A&A, 624, A134
Audren B., Lesgourgues J., Benabed K., Prunet S., 2013, J. Cosmol. Astropart. Phys., 2013, 001
Bartelmann M., Schneider P., 2001, Phys. Rep., 340, 291
Box G. E. P., Cox D. R., 1964, J. R. Stat. Soc. B, 26, 211
Charnock T., Battye R. A., Moss A., 2017, Phys. Rev. D, 95, 123535
Efstathiou G., Lemos P., 2018, MNRAS, 476, 151 (E18)
Feeney S. M., Peiris H. V., Williamson A. R., Nissanke S. M., Mortlock D. J., Alsing J., Scolnic D., 2019, Phys. Rev. Lett., 122, 061105
Gelman A., li Meng X., Stern H., 1996, Statistica Sinica, 6, 733
Grandis S., Rapetti D., Saro A., Mohr J. J., Dietrich J. P., 2016a, MNRAS, 463, 1416
Grandis S., Seehars S., Refregier A., Amara A., Nicola A., 2016b, J. Cosmol. Astropart. Phys., 2016, 034
Handley W., 2019, J. Open Source Softw., 4, 1414
Handley W., Lemos P., 2019a, Phys. Rev. D, 100, 023512
Handley W., Lemos P., 2019b, Phys. Rev. D, 100, 043504 (H19)
Heymans C. et al., 2012, MNRAS, 427, 146
Heymans C. et al., 2013, MNRAS, 432, 2433
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hildebrandt H. et al., 2020, A&A, 633, A69
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jenkins C. R., Peacock J. A., 2011, MNRAS, 413, 2895
Joachimi B., Taylor A. N., 2011, MNRAS, 416, 1010
Joudaki S. et al., 2017, MNRAS, 465, 2033
Kilbinger M., 2015, Rep. Prog. Phys., 78, 086901
Köhlinger F. et al., 2017, MNRAS, 471, 4412
Köhlinger F., Joachimi B., Asgari M., Viola M., Joudaki S., Tröster T., 2019, MNRAS, 484, 3126 (K19)
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Kunz M., Trotta R., Parkinson D. R., 2006, Phys. Rev. D, 74, 023503
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lin W., Ishak M., 2017a, Phys. Rev. D, 96, 023532
Lin W., Ishak M., 2017b, Phys. Rev. D, 96, 083532
Marshall P., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302
Nicola A., Amara A., Refregier A., 2019, J. Cosmol. Astropart. Phys., 2019, 011
Planck Collaboration XIII, 2015, A&A, 624, A134
Planck Collaboration, 2018, preprint (arXiv:1807.06209)
Raveri M., Hu W., 2019, Phys. Rev. D, 99, 043506
Rhodes J. et al., 2017, ApJS, 233, 21
Riess A. G. et al., 2018, ApJ, 855, 136
Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, ApJ, 876, 85
Schuhmann R. L., Joachimi B., Peiris H. V., 2016, MNRAS, 459, 1916
Schuhmann R. L., Heymans C., Zuntz J., 2019, preprint (arXiv:1901.08586)
Seehars S., Amara A., Refregier A., Paranjape A., Akeret J., 2014, Phys. Rev. D, 90, 023533
Troxel M. A. et al., 2018a, Phys. Rev. D, 98, 043528
Troxel M. A. et al., 2018b, MNRAS, 479, 4998
Woodbury M. A., 1950, Inverting Modified Matrices. Statistical Research Group, Memo. Rep. no. 42. Princeton Univ. Press, Princeton, NJ
Zhang F. Z., 2005 , The Schur Complement and Its Applications. Springer-Verlag, New York

APPENDIX

We use equation (7) to derive equation (9) [with the associated definitions in equation (10) through equation (13)] using the general 'complete the square' formula

$$(Q\theta - K)^{\mathsf{T}}\Sigma^{-1}(Q\theta - K) = (\theta - C)^{\mathsf{T}}(Q^{\mathsf{T}}\Sigma^{-1}Q)(\theta - C)$$
$$+ K^{\mathsf{T}}\Sigma^{-1}(K - QC), \tag{A1}$$

where $C = (Q^{\mathsf{T}}\Sigma^{-1}Q)^{-1}Q^{\mathsf{T}}\Sigma^{-1}K$. In equation (A1) set

$$Q = \begin{bmatrix} I \\ I \end{bmatrix}, K = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \text{ and } \Sigma = \Sigma_1 \text{ (so } \Sigma^{-1} = \Lambda). \tag{A2}$$

Also define $P = \begin{bmatrix} I & -I \end{bmatrix}$ (so, $PK = \mu_A - \mu_B$) and set

$$M^{-1} = \Sigma_A - \Sigma_X - \Sigma_X^{\mathsf{T}} + \Sigma_B = P\Sigma P^{\mathsf{T}} \tag{A3}$$

and

$$L^{-1} = \Lambda_A + \Lambda_X + \Lambda_X^{\mathsf{T}} + \Lambda_B = Q^{\mathsf{T}}\Lambda Q. \tag{A4}$$

The covariance matrix in the RHS of equation (A1) is

$$Q^{\mathsf{T}}\Sigma_1^{-1}Q = Q^{\mathsf{T}}\Lambda Q = L^{-1}; \tag{A5}$$

this will be Σ_0^{-1} in equation (9). For the constant term on the RHS of equation (A1), begin with

$$0 = PQ = (P\Sigma)(\Lambda Q) = \left(\Sigma_A - \Sigma_X^{\mathsf{T}}\right)\left(\Lambda_A + \Lambda_X\right)$$
$$+ \left(\Sigma_X - \Sigma_B\right)\left(\Lambda_X^{\mathsf{T}} + \Lambda_B\right), \text{ so}$$

$$\left(\Lambda_B + \Lambda_X^{\mathsf{T}}\right)\left(\Lambda_A + \Lambda_X\right)^{-1} = \left(\Sigma_B - \Sigma_X\right)^{-1}\left(\Sigma_A - \Sigma_X^{\mathsf{T}}\right)$$

$$L^{-1}(\Lambda_A + \Lambda_X)^{-1} - I = (\Sigma_B - \Sigma_X)^{-1}M^{-1} - I$$

$$(\Lambda_A + \Lambda_X)L = M(\Sigma_B - \Sigma_X). \tag{A6}$$

This establishes the upper-right block of the identity $\Lambda QLQ^{\mathsf{T}} + P^{\mathsf{T}}MP\Sigma_1 = I$ and the other blocks can be established similarly. Thus

$$K^{\mathsf{T}}\Sigma_1^{-1}(K - QC) = K^{\mathsf{T}}(I - \Lambda QLQ^{\mathsf{T}})\Lambda K$$
$$= (PK)^{\mathsf{T}}MPK = (\mu_A - \mu_B)^{\mathsf{T}}\Sigma_{\Delta\mu}^{-1}(\mu_A - \mu_B); \tag{A7}$$

this will contribute to $\log \mathcal{L}_0^{\max}$ in equation (9). Finally, the mean used in the RHS of equation (A1) is

$$C = (Q^{\mathsf{T}}\Sigma^{-1}Q)^{-1}Q^{\mathsf{T}}\Sigma^{-1}K = LQ^{\mathsf{T}}\Sigma_1^{-1}K = \Sigma_0 Q^{\mathsf{T}}\Lambda K$$

$$= \Sigma_0\left(\left(\Lambda_A + \Lambda_X^{\mathsf{T}}\right)\mu_A + (\Lambda_X + \Lambda_B)\mu_B\right); \tag{A8}$$

this will be μ_0 in equation (9).

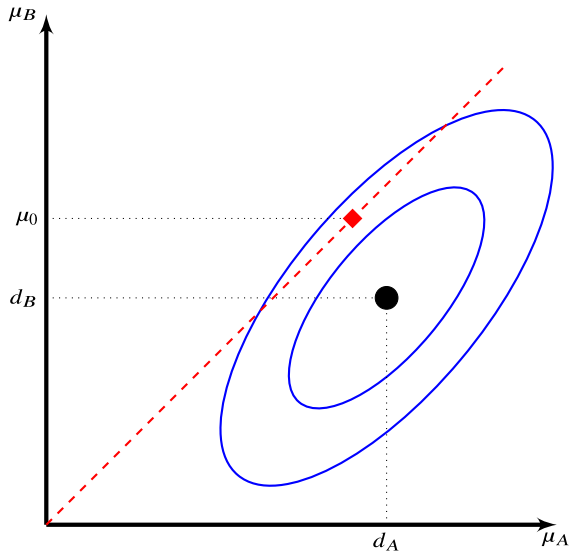See Fig. A1 for a diagram illustrating the behaviour of the likelihood function under the two hypotheses.

Figure A1. Inference of mean(s) from two correlated data observations d_A and d_B. The blue contours show (under hypothesis H_1, i.e. separate means) the likelihood (assumed Gaussian with covariance Σ) of the mean values (μ_A, μ_B); the maximum likelihood (and with flat priors the mode of the posterior) for these means is the black circle at (d_A, d_B). Under hypothesis H_0 (i.e. one common mean) the likelihood is restricted to the red dashed line $\mu_A = \mu_B$ where it will be Gaussian with covariance Σ_0; here the maximum likelihood (and hence posterior mode) is at the red diamond (μ_0, μ_0).

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