Debiasing inference with approximate covariance matrices and other unidentified biases

Elena Sellentin,[a] **Jean-Luc Starck**[b]

[a]Leiden Observatory, Leiden University, Huygens Laboratory, Niels Bohrweg 2, NL-2333 CA Leiden, The Netherlands.

[b]AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France.

E-mail: sellentin@strw.leidenuniv.nl

Abstract. When a posterior peaks in unexpected regions of parameter space, new physics has either been discovered, or a bias has not been identified yet. To tell these two cases apart is of paramount importance. We therefore present a method to indicate and mitigate unrecognized biases: Our method runs any pipeline with possibly unknown biases on both simulations and real data. It computes the coverage probability of posteriors, which measures whether posterior *volume* is a faithful representation of *probability* or not. If found to be necessary, the posterior is then corrected. This is a non-parametric debiasing procedure which complies with objective Bayesian inference.

We use the method to debias inference with approximate covariance matrices and redshift uncertainties. We demonstrate why approximate covariance matrices bias physical constraints, and how this bias can be mitigated. We show that for a Euclid-like survey, if a traditional likelihood exists, then 25 end-to-end simulations suffice to guarantee that the figure of merit deteriorates maximally by 22 percent, or by 10 percent for 225 simulations. Thus, even a pessimistic analysis of Euclid-like data will still constitute an 25-fold increase in precision on the dark energy parameters in comparison to the state of the art (2018) set by KiDS and DES. We provide a public code of our method.

Contents

1 Introduction: unrecognized biases or new physics?

The hardest mistakes to correct for, are those which remained unnoticed, or for which no solution exists yet. Contemporary cosmology actively tackles biases from covariance matrices [1–4], likelihoods [5, 6], lacking spectroscopic data for redshifts [7, 8], and [9] lists a comprehensive review of many more difficulties in leading weak lensing [10] data analyses.

Known and unknown biases propagate into cosmological parameter constraints, where they cause shifts of the posterior. In the absence of any biases, a posterior peaking in unexpected regions of parameter space must however be interpreted as a sign of new physics, and it is therefore of utmost importance to tell unrecognized biases and new physics apart. Furthermore, this distinction needs to be convincing beyond the boundaries of cosmology, i.e. also be convincing for neighbouring fields such as particle physics.

We therefore here provide a method which safeguards cosmological parameter constraints against recognized or unrecognized biases.

Based on a joint analysis of simulations and the real data with a likelihood, the method leads to unbiased credibility contours for the physical parameters. The method is non-Bayesian (but compatible with Bayesian inference) and therefore applies also when there is no error model available, which a Bayesian mitigation method would require. The thus gained credibility contours have a precise mathematical meaning, namely that of correct 'coverage probability' (defined in section 2). Coverage probabilities of Bayesian posteriors objectively measure differences between frequentist and Bayesian parameter constraints. They thereby measure how much the inferred physics depends on our *assumptions* when analyzing the data, rather than on information contained *in* the data. Accordingly, reporting the coverage also measures how much (frequentist) particle physicists, and (Bayesian) cosmologists could maximally disagree, given the same data set.

We develop our method in section 2. The method is general, but was developed to address currently outstanding problems of cosmic shear. For example, [1, 2] derive the to-date only known completely bias-free likelihood for estimated covariance matrices. In [11], it was then shown that extra-correlations exist between weak lensing data points, which cannot be captured by any covariance matrix, but affect the inference. In [5], these extra-correlations were studied in detail, showing that the actual weak lensing likelihood must be skewed, and that this skewness translates into parameter biases up to 10 percent of the standard deviation, depending on how the weak lensing data are binned in angular ranges and redshift bins.

A recurrent theme in these analyses was that weak lensing does not easily [12] lend itself to simulations, due to reacting to cosmic structures on the scale of galaxy groups, and due to these structures falling already into the strongly non-linear regime of structure formation. We therefore here seek to minimize the *number* of simulations, thereby trading for high accuracy of the few simulations, and nonetheless gaining faithful parameter constraints from a joint analysis of data and simulations with a likelihood.

Section 3.4 mitigates parameter biases from approximate likelihoods, where our example uses approximate covariance matrices. Section 3.5 studies photometric redshift uncertaintites and shows that redshift uncertaintites alone (without biased redshifts) can be neglected in current weak lensing surveys. Section 4 shows that 25 end-to-end simulations of a Euclid-like [13] survey, in conjunction with an independent likelihood for this survey, suffice to guarantee that the figure of merit deviates maximally by 22 percent from its optimum. For 225 end-to-end simulations, the figure of merit can be guaranteed to deteriorate by maximally 10 percent. As a result, it can be taken essentially for granted that the upcoming Euclid-like surveys will lead to an 25-fold increase in our knowledge of the dark energy equations of state parameters [14] w_0 and w_a.

2 Mitigating unrecognized biases: method and examples

To avoid that unrecognized biases feign new physics, we establish a method that takes as input any existing data analysis pipeline. The method runs the pipeline on simulations and real data alike, and then computes and corrects the *coverage probability*. We describe why this procedure debiases parameter constraints.

2.1 What do posteriors really measure?

Biases in an inference cause that a posterior, or likelihood, exclude the true parameters too often, for example because the posterior is shifted or too narrow. The notion of 'too often', is made mathematically precise by *coverage probabilities*. The coverage probability of a posterior credibility contour is the fraction of times that this contour includes the true parameters, under repetitions of the experiment. The default expectation of most scientists is that the 68 percent credibility contour (as an example) contains the true parameters 68 percent of the time. In reality, however, the 68 percent posterior credibility contour is constructed such that it contains 68 percent of the *posterior volume*. Most scientists expect that posterior volume measures (Kolmogorov) probability, but this is not necessarily so. We refer to this expectation by speaking of 'correct coverage' for short [15, 16].

Coverage probabilities superficially sound like a frequentist concept, but so-called objective Bayesian analyses [17] achieve the correct coverage probabilities as well [18, 19], due to their explicit construction of priors. Objective Bayesian analyses thus implement the correct noise propagation through mathematical derivations, with the result that posterior volume

indeed measures probability. In contrast, the correct coverage is not automatically achieved by so-called subjective Bayesian analyses [20, 21]. These regard priors as subject to choice, or use hyper-parameters, approximate likelihoods [5, 11], or idealized parametric models, with the result that the total Bayesian flow of information is not representative of nature, although mathematically self-consistent [15, 19].

In total, it cannot be taken for granted that posterior volume measures probability as expected, but such potential discrepancies can be *reported* by quoting coverage probabilities. This is of direct relevance to tensions between experiments.

2.2 Algorithm to measure the coverage probability of posteriors

Any unrecognised or unintended systematic will affect the coverage. Hence, measuring the coverage can detect hidden biases, even if the source of the biases is unknown. Correcting the coverage is then a model-independent solution for mitigating unrecognized biases. We measure and correct the coverage probability with the following algorithm.

A set of fiducial parameters $\boldsymbol{\theta}_0$ is chosen for N simulations of artificial data sets $\boldsymbol{x}_i$, with $i \in [1, N]$. These simulations imitate the real data $\boldsymbol{x}_o$. We denote posterior *densities* by curly capital $\mathcal{P}$, and associated probabilities, that are scalar rather than densities, by roman P.

A state-of-the-art likelihood is then run on all simulations and also on the real data. This results in N posteriors $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}_i)$ from simulations, and the posterior $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}_o)$ of the real data. For each of these $N + 1$ posteriors, 120 credibility contours (or more) are computed. We provide a public code[1], where 100 of these contours are equidistant between zero and 99.9 percent posterior credibility. Twenty further contours are equidistant between 95.25 and 99.75 percent credibility. These finely spaced contours enable a reliable coverage correction in the outer tails of a posterior. If the data analysis pipeline contains biases, then the contours resulting from it will not cover correctly.

We denote by α fractions of the posterior volume, and accordingly $\alpha \in [0, 1]$. We consider posterior contours that contain a fraction α of the posterior volume and which are isocontours of the posterior. They thus cut the posterior in a certain height below its peak. For each data set i, the posterior will be slightly differently shaped, and the height of the αth contour thus changes with i. We therefore denote this height as roman P_α^i, where $i \in [0, N]$ identifies the data set, and $\alpha \in [0, 1]$ identifies the fraction of posterior volume that the contour contains.

Each of the posteriors will take a different (scalar) value *at* the fiducial parameters of the simulation. We denote this value as roman $P(\boldsymbol{\theta}_0|\boldsymbol{x}_i)$, where the subscript zero indicates that this is the posterior probability assigned to the fiducial parameters.

The αth credibility contour then contains the true parameters if the posterior value at the fiducial parameters is larger than the posterior height of the contour:

$$P(\boldsymbol{\theta}_0|\boldsymbol{x}_i) \geq P_\alpha^i \Rightarrow \ \alpha\text{th contour contains } \boldsymbol{\theta}_0. \tag{2.1}$$

We measure this for all contours, for all posteriors. The coverage probability, C_α, is then the probability p that the α-posterior credibility region contains ('covers') the true parameter values

$$C_\alpha = p(\boldsymbol{\theta}_0 \text{ inside } \alpha\text{th contour}). \tag{2.2}$$

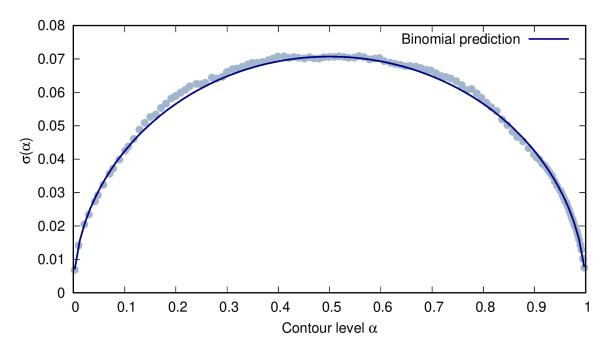[1]Public at github.com/elenasellentin/Mitigate_Unrecognized_Biases

Figure 1: Standard deviation of the coverage estimator $\hat{C}_\alpha$, as a function of the contour level α. The grey points indicate the numerically estimated standard deviations of a coverage measured from N simulations, and the blue solid line indicates the prediction for the standard deviation from a binomial distribution.

The default expectation would be that $C_\alpha = \alpha$, meaning that posterior volume measures probability under repetition of the experiment. In contrast, if biases occurred in the analysis, then a credibility contour further out in the posterior will achieve coverage C_α.

For example, the allegedly 95 percent credibility contour of the biased analysis might be found to contain the true parameters only 90 percent of the times. Then it is in reality the 90 percent contour, until the bias is found and corrected. If the bias cannot be found, a mitigation is to discard the biased contours and instead adopt the contours of correct coverage. The new contours will then include the true parameters with the right fraction of times – despite the bias being unknown.

The coverage of equation 2.2 can be estimated from N simulations, and we denote its estimator by $\hat{C}_\alpha$. This estimator simply counts how often the true parameters fall *inside* the α-contour. If they do not fall inside the contour, they fall automatically outside, and this either-or process indicates that the estimator $\hat{C}_\alpha$ must (by definition) follow a binomial distribution with success rate α and N trials. The mean and standard deviation of the binomial distribution then give the mean and standard deviation σ of our coverage estimator

$$\langle \hat{C}_\alpha \rangle = C_\alpha, \quad \sigma = \frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{N}}. \tag{2.3}$$

Figure 1 shows that the binomial distribution models the noise in the estimated coverage correctly: for the innermost contours, where α is low, and for the outermost contours, the standard deviation is the smallest.

If credibility contours cover correctly, then $C_\alpha = \alpha$, and the standard deviations will in the following be adopted as error bars.
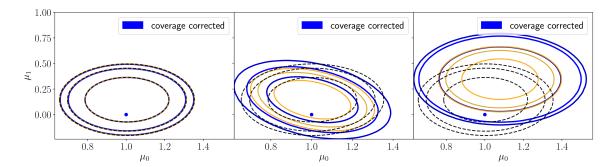
Figure 2: Examples of unrecognized biases corrected via coverage measurements. Dashed contours: the correct, bias-free analysis for comparison. Yellow: biased analysis per panel (left: no bias, middle: bias introduced by an approximate covariance matrix, right: bias introduced by an unintentionally informative prior). Blue dot: the true parameters, note how the yellow posterior in the right is so biased that it excludes this point. Blue contours: coverage corrected analysis, the true parameter point is now included with the right probability, even if the bias cannot be found.

We provide three simple examples of coverage correction in section 2.3, before we apply the method to cosmological analyses in section 3.

2.3 Examples

Figure 2 illustrates three examples of mitigating unrecognized biases via coverage correction. The real data vector contains 100 data points, drawn from a Gaussian distribution with unit variance. The first fifty data points have mean $\mu_0 = 1$, the remaining data points have $\mu_1 = 0$. The parameters to be inferred are $p_0 = \hat{\mu}_0$ and $p_1 = \hat{\mu}_1$. We simulate 1000 artificial data vectors, by drawing from the same Gaussian. A bias is then introduced in the analysis, the coverage is measured and corrected, resulting in increased contour size.

Example 1, in the left panel of figure 2 is bias free: the data are analyzed with the correct Gaussian likelihood and a flat unbounded prior, which produces automatically the correct coverage for linear parameters, such as μ_1, μ_2. Example 2, in the middle panel, analyzes the data with a biased inverse covariance matrix. The correct inverse covariance would have been $\mathsf{C}^{-1} = \mathbb{I}$, the identity matrix, but the off-diagonal elements were changed to $C_{ij}^{-1} = 1e^{-2}$. Section 3 will detail why biased covariance matrices shift posteriors, here we only illustrate that our method is able to correct for this, without needing to know the origin of the error.

Figure 3 plots the measured coverage probabilities of example 2. Due to the hidden bias, the contours are systematically to small, resulting in the seen undercoverage of figure 3.

Finally, example 3 corrects the effects of an unintentionally informative prior π, given by

$$\pi(p_0, p_1) = \mathcal{G}(p_0|\mu = 1.45, \sigma^2 = 0.1)\mathcal{G}(p_1|\mu = 1.35, \sigma^2 = 0.1), \qquad (2.4)$$

where $\mathcal{G}$ is the Gaussian distribution. The prior is so informative that the biased posterior excludes the true parameters (blue point). After coverage correction, the true parameters are again included. Plotted contours lie at 68, 90 and 95 percent posterior volume (before coverage correction, yellow), and at 68, 90 and 95 percent coverage probability (after correction, blue).
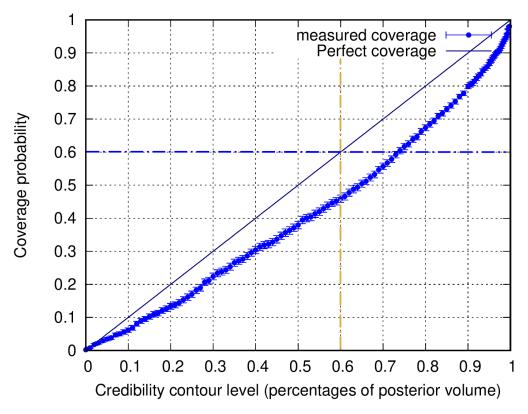
Figure 3: Measured coverage probabilities of example 2, the middle panel of figure 2, where the bias is due to an approximate covariance matrix. Horizontal axis: the credibility contours, computed as percent of (biased) posterior volume, where the posterior is too narrow. The vertical axis shows the coverage probabilities of the contours, this is the probability that a given contour includes the true parameter values. The diagonal line is a bias-free analysis, where posterior volume reflects probability. However, due to the bias of example 2, all contours contain the true parameters systematically fewer times than expected (they are too narrow). For example, the 60% credibility contour contains 60% of posterior *volume* (yellow cut), but it contains the true parameter point only 48% of the times. After measuring the coverage, this can be corrected: the blue cut shows that the (biased) 72% posterior volume contour contains the true parameters 60% of the time. The (biased) 72% volume contour is thus the (unbiased) 60% confidence contour. The unknown bias is thus mitigated by relabelling the contours.

2.4 Blind spots of the method

The method detects discrepancies between simulations and the assumptions of a data analysis pipeline. It then corrects for these discrepancies when analyzing the real data. Consequently, it cannot correct for systematic effects which are omitted in both simulations and the analysis pipeline. For example, if neither a likelihood, nor the simulations include a survey mask, then the method cannot correct for imperfections in survey mask handling. Likewise, if the simulations implement precisely the same assumptions as the analysis pipeline, then a self-confirming situation is created, which the method also cannot detect. If the posterior then peaks nonetheless in unexpected regions, then the real data obey other physical or statistical

laws than the ones simulated.

If the simulations lack in accuracy, then our method suffers from the same difficulties as any simulation-based inference technique. We shall however find in section 4 that our method requires by many orders of magnitude the fewest number of simulations [2, 3]. This arises due to the the joint analysis with a likelihood, which already contains statistical information which simulations would otherwise need to provide.

3 Applications to cosmology

In this section we apply coverage calibration to cosmic shear (weak lensing) analyses [7, 8, 10], where approximate covariance matrices and redshift uncertaintites often introduce biases of unknown magnitude and of unknown parametric form.

3.1 Why approximate covariance matrices shift posteriors

Approximate covariance matrices are today used in weak lensing [7, 8], but also supernova analyses adapt their covariance matrices to achieve a desired goodness of fit [22, 23]. One often encountered preconception is that such approximate covariance matrices only affect the width of posterior contours, but not where a posterior peaks. We therefore explain why the opposite is true: We show that using an approximate covariance matrix is mathematically the same as fitting to a biased data set, and systematic parameter shifts will ensue.

Consider a Gaussian likelihood, as is currently standard in cosmology

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \mathsf{C}) = \frac{1}{\sqrt{(2\pi)^p |\mathsf{C}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \mathsf{C}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{3.1}$$

where, $\boldsymbol{x}$ is a p-dimensional data vector and the superscript $\top$ denotes transposition. The mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a function of the parameters $\boldsymbol{\theta}$ to be inferred, and the covariance matrix is C. Parameters are then estimated by sampling the posterior, which is the likelihood times a prior.

To isolate the effect of approximate covariance matrices, we assume unbounded flat priors, and that the data $\boldsymbol{x}$ contain no systematic effects. Such sound data are then nonetheless *effectively* transformed into a biased data set, if an approximate covariance matrix is used in the analysis. This can be seen as follows.

Let the correct covariance matrix be C_c and let C_B be an approximation of it. Both are symmetric positive-definite matrices.

If an analysis uses the correct covariance matrix, the best fit lies where

$$\chi_c^2 = (\boldsymbol{x} - \boldsymbol{\mu}_c)^\top \mathsf{C}_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c), \tag{3.2}$$

is minimal. Equation 3.2 describes that the parameters of the model $\boldsymbol{\mu}$ will adjust to minimize the distance to the data $\boldsymbol{x}$. The best-fitting parameters are then $\boldsymbol{\theta}_c$ for which $\boldsymbol{\mu}_c = \boldsymbol{\mu}(\boldsymbol{\theta}_c)$. During minimization, statistical compatibility between the mean and the data is measured in units of the inverse covariance matrix. If we exchange the covariance matrix, this distance measure changes. In the units of the new covariance matrix, another $\boldsymbol{\mu}(\boldsymbol{\theta})$ will then be closest to the data $\boldsymbol{x}$. Consequently, the parameters $\boldsymbol{\theta}$ will adapt, in order to produce this new mean as well as possible.

We now relate the two matrices via the function

$$\mathsf{C}_B^{-1} = \mathsf{B}^\top \mathsf{C}_c^{-1} \mathsf{B}, \tag{3.3}$$

where a bias occurs if $\mathsf{B} \neq \mathbb{I}$. The left- and right-multiplication by B is convenient, but not a specialization. We could equally have written

$$\mathsf{C}_{\mathrm{B}}^{-1} = \mathsf{C}_{\mathrm{c}}^{-1} + \Delta, \tag{3.4}$$

where Δ is the matrix of additive inaccuracies. Since $\mathsf{C}_{\mathrm{c}}^{-1}$ is a symmetric matrix, Δ is also by construction symmetric. The matrix B is then guaranteed to exist, since for symmetric matrices A any congruent matrix $\mathsf{B}^{\top}\mathsf{A}\mathsf{B}$ is again symmetric for all B, and equations 3.3 and 3.4 are both valid ways of describing the systematic uncertainties in a covariance matrix. The corresponding additive uncertainty is then

$$\Delta = \mathsf{B}^{\top}\mathsf{C}_{\mathrm{c}}^{-1}\mathsf{B} - \mathsf{C}_{\mathrm{c}}^{-1}. \tag{3.5}$$

Since $\mathsf{C}_{\mathrm{c}}^{-1}$ is unknown, cosmology is forced to use $\mathsf{C}_{\mathrm{B}}^{-1}$ for the likelihood. The thus gained χ^2-squared surface is then minimized where

$$\chi_{\mathrm{B}}^2 = (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{B}})^{\top}(\mathsf{B}^{\top}\mathsf{C}_{\mathrm{c}}^{-1}\mathsf{B})(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{B}}), \tag{3.6}$$

is minimized. This occurs at a new mean $\boldsymbol{\mu}_{\mathrm{B}} = \boldsymbol{\mu}(\boldsymbol{\theta}_{\mathrm{B}})$, and the parameters $\boldsymbol{\theta}_{\mathrm{B}}$ will differ from $\boldsymbol{\theta}_{\mathrm{c}}$.

If we conduct a thought experiment where we forget that the new parameters differ, we see that using a biased covariance matrix is akin to analyzing a biased data set $\boldsymbol{x}_{\mathrm{B}}$ with the correct covariance matrix. To see this, we set $\boldsymbol{\mu}_{\mathrm{B}} = \boldsymbol{\mu}_{\mathrm{c}}$ in our thought experiment. Then, to yield as good a best fit as when using the correct covariance matrix, we have to demand

$$\mathsf{B}(\boldsymbol{x}_{\mathrm{B}} - \boldsymbol{\mu}_{\mathrm{c}}) = (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{c}}). \tag{3.7}$$

This can be solved for $\boldsymbol{x}_{\mathrm{B}}$, and we find

$$\boldsymbol{x}_{\mathrm{B}} = \mathsf{B}^{-1}\left[\boldsymbol{x} + \boldsymbol{\mu}_{\mathrm{c}}(\mathsf{B} - \mathbb{I})\right]. \tag{3.8}$$

This shows that using an incorrect covariance matrix $\mathsf{B}^{\top}\mathsf{C}_{\mathrm{c}}^{-1}\mathsf{B}$ to analyze a sound data set is mathematically equivalent to analyzing the biased data set $\boldsymbol{x}_B$ with the correct covariance matrix. Only if B equals the identity matrix does $\boldsymbol{x}_{\mathrm{B}}$ coincide with $\boldsymbol{x}$.

In cosmology, the data $\boldsymbol{x}$ are of course fixed. The only free variables to compensate for the bias in the covariance matrix are then the cosmological parameters $\boldsymbol{\theta}$. The incorrect covariance matrix will consequently force the likelihood to peak at biased parameter values.

In fact, in order for the biased equation 3.6 to reproduce as good a fit as the correct equation 3.2, the relation

$$\mathsf{B}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{B}}) = \boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{c}}, \tag{3.9}$$

needs to hold. Solving for the now preferred $\boldsymbol{\mu}_{\mathrm{B}}$, we find

$$\boldsymbol{\mu}_{\mathrm{B}} = \boldsymbol{x} - \mathsf{B}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{c}}). \tag{3.10}$$

The parameters $\boldsymbol{\theta}$ will then attempt to create the mean $\boldsymbol{\mu}_{\mathrm{B}}$ instead of the mean $\boldsymbol{\mu}_{\mathrm{c}}$. Depending on the flexibility of the model, the parameters may not fully succeed in this. In total, we see however that a shift in parameters will ensue, and the direction and magnitude of the shift depends on the drawn data vector $\boldsymbol{x}$, and the biasing matrix B, according to equation 3.10.

3.2 Undetectability in Fisher matrix forecasts

The effect of approximate covariance matrices biasing parameters is invisible in Fisher matrix forecasts [13, 24], because on average, one has $\langle \boldsymbol{x} \rangle = \boldsymbol{\mu}_{\mathrm{c}}$, and equation 3.10 then predicts $\boldsymbol{\mu}_{\mathrm{B}} = \boldsymbol{\mu}_{\mathrm{c}}$. Fisher forecasts will therefore underestimate the total uncertainty. The biasing effect will only occur when analysing real data, where $\boldsymbol{x}$ is fixed to the realization on the sky. Equation 3.10 then describes that noise can be incorrectly interpreted as 'signal' when the wrong covariance matrix is employed. In the following section we will present an example of thus resulting parameter biases.

3.3 Forcing the KiDS-450 data to prefer the Planck cosmology

Concerning how approximate covariance matrices bias physical parameters, we here illustrate that direction and magnitude of the posterior shift can also be controlled. Additionally, the goodness of fit can also be kept constant. A reduced-χ^2 of order unity is therefore by no means a good indication that the best-fitting parameters are unbiased.

We illustrate this for the public KiDS-450 data from [7], and force these data to prefer the Planck cosmology. DES [8] analyses could equally have been used. We underline that we here *force* this transition to the Planck best-fitting cosmology. The aim of this study is dual, namely first to understand which data points are affected, and secondly to understand which procedures must be put in place in order to prevent such shifts.

We work with the original KiDS-450 data vector of 130 elements, which are the real-space estimators ξ_+ and ξ_- [10, 25] in four tomographic redshift bins and their cross-bins. Our weak lensing setup to compute the theory vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ is identical to [7] and our code has been verified against the code of [7], leading to identical results for the theory vectors, given identical input parameters. We use CLASS [26, 27], and HALOFIT [28] for the non-linear power spectrum. We fix the spectral index n_s and the reduced Hubble constant h, to Planck-motivated values of $h = 0.678$ and $n_s = 0.96$. Varying the cold dark matter density Ω_m and the normalization of the power spectrum σ_8, we find the best-fitting cosmology for the KiDS-450 data vector when analyzed with the public KiDS-450 covariance matrix to be

$$\Omega_m = 0.2, \ \sigma_8 = 0.838, \tag{3.11}$$

with a $\chi^2 = 202$[2]. By transforming the covariance matrix, we now force the KiDS-450 data to prefer the Planck cosmology. This can be repeated for arbitrarily many parameters.

We precompute the cosmological predictions $\boldsymbol{\mu}(\boldsymbol{\theta})$ on a grid, and store the results, in order to make the upcoming analyses of this paper numerically feasible.

We use a Planck best-fitting cosmology with [29]

$$\begin{aligned} n_s = 0.96, \ h = 0.678 \\ \Omega_m = 0.308, \ \sigma_8 = 0.83. \end{aligned} \tag{3.12}$$

The original KiDS-450 analysis [7] leads to posterior constraints on σ_8 and Ω_m which are in tension with the Planck constraints. The left panel of figure 4 plots the result of subtracting the KiDS-best fitting cosmology, or the Planck best-fitting cosmology from the KiDS data vector. Subtracting the Planck best-fitting cosmology leads to multiple sequences of adjacent data points being systematically below the mean (blue triangles in the negative domain). A

---

[2]The high value of this χ^2 results from having fixed n_s and h to the Planck best-fitting values, rather than the KiDS best-fitting values.
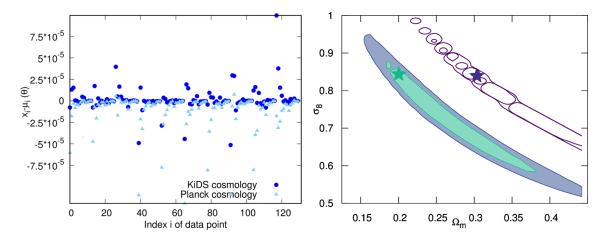
Figure 4: Left: The differences $\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})$ for the KiDS best-fitting cosmology (blue dots), and the Planck best-fitting cosmology (triangles). To achieve a translation of the posterior from the original KiDS best-fit (located at the green star in the right plot) to the Planck best-fit (located at the purple star in the right plot), a covariance matrix must be constructed which expects the noise pattern on the left as an indication of 'strong covariance'. Right: The posterior with solid contours uses the original KiDS data and the original KiDS covariance matrix. The posterior in open contours uses the original KiDS data and the newly constructed covariance matrix. Its deformed shape results from having changed the determinant of the covariance matrix. The contours lie at 68% and 90% of posterior volume (which is the standard procedure in cosmology).

Table 1: Data points of KiDS-450 which are most unstable with respect to noise reassessment. The data point identifier '#' counts from 1 and has the KiDS-ordering. The angular cuts in the corresponding DES data [8] exclude these data points. Together with [11], the effect now repeatedly occured that data vector truncation influences the physical parameter constraints, which motivates that blinding strategies should be kept for future analyses.

#	angle (arcmin)	$\xi_\pm$	#	angle (arcmin)	$\xi_\pm$
66	0.713	ξ_+	92	0.713	ξ_+
40	0.713	ξ_+	53	0.713	ξ_+
105	0.713	ξ_+	79	0.713	ξ_+
1	0.713	ξ_+	14	0.713	ξ_+
42	2.956	ξ_+	54	1.452	ξ_+
67	1.452	ξ_+	80	1.452	ξ_+
106	1.452	ξ_+	107	2.956	ξ_+
119	1.452	ξ_+	120	2.956	ξ_+
95	6.017	ξ_+	16	2.956	ξ_+
28	1.452	ξ_+	41	1.452	ξ_+

covariance matrix can be tricked into expecting such a situation: By definition we have that the covariance between data points x_i and x_j is

$$c_{ij} = \mathbb{E}(x_i x_j) - \mathbb{E}(x_i)\mathbb{E}(x_j), \tag{3.13}$$

where $\mathbb{E}$ denotes taking the expectation value. Since this is an expectation value, a covariance
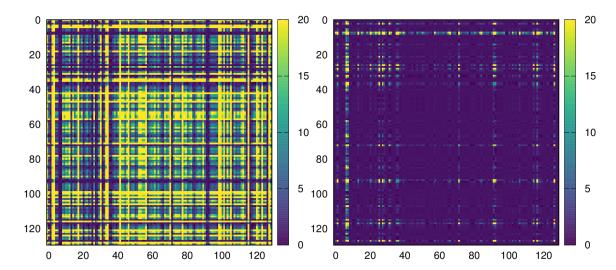
Figure 5: Left: Relative difference between the KiDS covariance matrix C, and the transformed matrix $\mathsf{B}^{-1}\mathsf{C}\mathsf{B}^{-1}$. Right: As left, but now for the inverse covariance matrix. Changes in only a few columns of the precision matrix are sufficient to induce major biases in the physical parameters.

matrix does not simply describe noise, but is rather extremely prescriptive: a positive covariance between data point x_i and x_j describes that *if* data point x_i is below the mean, then data point x_j is *expected* to be below the mean as well. We can hence construct a covariance matrix that expects the noise pattern of the blue triangles in figure 4 and interprets it as a strong positive correlation between all data points that are below the mean. The data points whose noise will thereby be most strongly reassessed are listed in table 1, which illustrates that it is consistently the estimators ξ_+ on the lowest angular scales (mostly 0.71 and 1.45 arcmins) who will show instability with respect to cosmological parameters, when their noise is reassessed. In this context it is interesting to note that the DES survey [8] excludes ξ_+ on such low scales, which will be partially responsible for why DES posteriors are closer to Planck than KiDS-450 posteriors.

We denote the original KiDS-450 covariance matrix as $\mathsf{C}_{\mathrm{KiDS}}$. The minimal χ^2 is then reached for

$$\chi^2 = 202 \text{ for } \mathsf{C}_{\mathrm{KiDS}}, \text{ at } \boldsymbol{\theta}_{\mathrm{c}} = (0.2, 0.838, 0.678, 0.96), \tag{3.14}$$

where the parameter vector is ordered as $\boldsymbol{\theta} = (\Omega_m, \sigma_8, h, n_s)$. We now demand that the KiDS data vector instead produces the Planck cosmology $\boldsymbol{\theta}_{\mathrm{B}} = (0.308, 0.83, 0.678, 0.96)$ as best fit, and solve for the matrix B from equation 3.10 which enables this.

Since the matrix B has $p \times p$ entries, but equation 3.10 only poses p constraints, reconstructing B is an under-determined system. There will hence be infinitely many solutions for B, which directly implies that trying to debias an approximate covariance matrix is bound to fail.

Here, we now pick out one solution, by demanding B to be diagonal, $\mathsf{B} = \mathrm{diag}(b_1, b_2, ..., b_p)$. The required diagonal elements to force the KiDS data to prefer the Planck cosmology then follow to be

$$b_n = \frac{(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{c}})_n}{(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{B}})_n} \quad \forall n \in [1, p]. \tag{3.15}$$

Using the original KiDS-450 data vector, and transforming the inverse KiDS covariance matrix to $\mathsf{C}_{\mathrm{KiDS}}^{-1} \to \mathsf{B}^\top \mathsf{C}_{\mathrm{KiDS}}^{-1} \mathsf{B}$, the Planck cosmology indeed becomes the new best fit

$$\text{KiDS data, } \mathsf{B}^\top \mathsf{C}_{\mathrm{KiDS}}^{-1} \mathsf{B} : \Delta\chi^2 = 0 \text{ at } \boldsymbol{\theta}_{\mathrm{Planck}}. \tag{3.16}$$

The two posteriors arising from analyzing the KiDS data with the two covariance matrices are depicted in figure 4. This figure illustrates the successful translation of the posterior, although data and physical model were not changed. The Planck cosmology now fits the KiDS data with the same goddness of fit (the same χ^2) as the KiDS best-fitting cosmology fitted the KiDS data before. Also visible is, however, that the new posterior is deformed. This side effect arises because the determinant of the covariance matrix was changed[3].

We compute the relative differences between the original and the transformed covariance matrices. The matrix of relative differences is given by

$$R_{ij} = \frac{|C_{ij} - \tilde{C}_{ij}|}{|C_{ij}|}, \tag{3.17}$$

where i and j are the indices of the matrix elements and $\tilde{\mathsf{C}}$ is shorthand for the transformed matrices. The left panel of figure 5 shows the relative difference matrix R_{ij} for the transformed covariance matrix $(\mathsf{B}^{-1})^\top \mathsf{C}_{\mathrm{KiDS}} \mathsf{B}^{-1}$, and the right panel shows the relative difference matrix R_{ij} of the transformed inverse covariance matrix $\mathsf{B}^\top \mathsf{C}_{\mathrm{KiDS}}^{-1} \mathsf{B}$. The difference between the left and the right panel highlights the unpredictability of the matrix inversion: even if most columns in $\mathsf{C}_{\mathrm{KiDS}}$ are drastically changed, these changes can be redistributed during the inversion, and it is thus important to judge the accuracy of an *inverse* covariance matrix directly.

Figure 5 reveals factor 20 changes in certain elements of the inverse covariance matrix. This is to be compared to the DES reanalysis [30] of KiDS-450, where the reanalysis implemented factor 3 changes in the shape noise contribution to elements of the covariance matrix, and parameter shifts were found. We therefore conclude that a debiasing procedure for approximate covariance matrices is indeed needed.

As infinitely many solutions exist to induce a bias such that any arbitrary cosmology becomes the best-fitting cosmology, a parametric Bayesian treatment will not be able to debias the inference. In the following section we will hence reverse the workflow, and accept that any fixed covariance matrix of unknown bias will necessarily be used, and we debias the thus resulting parameter inference with coverage calibration.

3.4 Debiasing inference with approximate covariance matrices of unknown bias

In this section, we illustrate how to compute unbiased credibility contours for cosmological parameters, despite using a covariance matrix of unknown but non-zero bias.

A necessary prerequisite for our method are N accurate simulations. Importantly, these N simulations are not used to compute a covariance matrix, or its inverse – they are used to debias the inference pipeline which uses the approximate, analytical covariance matrix. To compute a numerical covariance matrix from simulations, one would require $N \gg p$, where p is the dimension of the data set. To run our debiasing procedure, significantly fewer

---

[3]Keeping the determinant constant would impose only one additional constraint, still leading to infinitely many solutions for B, again leading to the conclusion that a non-parametric method is needed to debias inference with approximate covariance matrices.
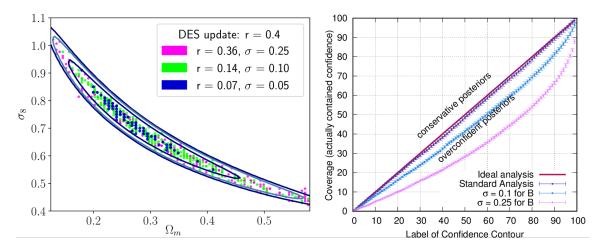
Figure 6: Left: The original KiDS posterior is depicted in open contours. The dots indicate how the best-fit scatters around if the KiDS data set is analyzed with differently biased covariance matrices. The entire posteriors shift along with the new best-fit, but the shifted contours are not shown for reasons of plot overcrowding. The relative biases ($r = 0.07, 0.14, 0.36$) here introduced to the covariance matrix are smaller than the changes applied by the DES reanalysis of KiDS ($r = 0.4$) [30]. Right: Measured coverages of the biased posteriors. The more biased the inverse covariance matrix is, the more does the biased posterior undercover (meaning it is too narrow). The posterior is debiased in figure 7.

simulations are needed, and their number does not scale with the dimension of the data set either, see equation 2.3 and section 4.

We again use KiDS-450 as an example. For current weak lensing studies, sufficiently many or accurate simulations do not yet exist to conduct a coverage measurement. KiDS-450 posesses 930 simulations for 100 square degree sky patches [31], but spans by itself approximately 450 square degree. DES uses 18 simulations in [32], where the number of simulations is now the limiting factor.

To demonstrate our method, we therefore generate 100.000[4] Gaussian realizations of data vectors with the KiDS best-fitting cosmology as mean, and with the public KiDS covariance matrix. These shall serve as our simulations replacement.

We then precompute the theory vectors $\boldsymbol{\mu}(\boldsymbol{\theta})$ on a grid in the Ω_m, σ_8-plane, and then compute the 100.000 posteriors. The posterior per data vector $\boldsymbol{x}$ is

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) \propto L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{3.18}$$

where $L(\boldsymbol{x}|\boldsymbol{\theta})$ is the Gaussian likelihood, and π are priors on the parameters. We use top-hat priors, with

$$0.09 < \Omega_m < 0.65, \quad 0.37 < \sigma_8 < 1.1. \tag{3.19}$$

Finally, the coverage is computed.

The coverage resulting from this analysis pipeline is plotted in the left of figure 6. The red diagonal line indicates the perfect coverage for an unbiased analysis. Measured coverage probabilities above the red line indicate conservative credibility contours, which are strictly

---

[4]This large number resulted from experimenting with run time constraints. Far fewer are needed in reality, see section 4 for Euclid requirements.
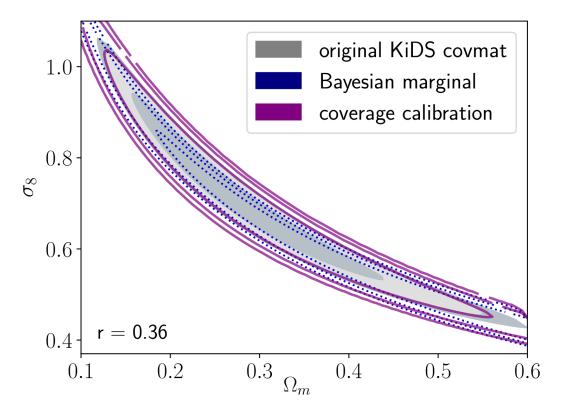
Figure 7: KiDS posteriors, with and without propagation of covariance matrix uncertainty. This plot uses $\sigma = 0.25$ which leads to relative changes of $r = 0.36$ in the covariance matrix. Given [30], these are realistic values for current covariance matrix uncertainties. Solid contours: original KiDS-450 posterior, without propagating covariance matrix uncertainty. Blue dotted: Propagating the uncertainty via Bayesian marginalization, here possible since the toy-model for the bias is known, but impossible in reality where the model is unknown. Purple open contours: debiasing the credibility contours via coverage calibration. Comparison of solid grey and purple contours: for $r = 0.36$, the 68% credibility contour of the debiased posterior is as large as the 90% posterior credibility contour of the biased posterior (grey).

speaking too wide. Measured coverage probabilities below the red line indicate credibility contours which are too narrow. The purple data points depict the measured coverage with error bars. As can be seen, the posterior with the correct covariance matrix undercovers slightly, meaning it is slightly too narrow. This reflects that the adopted priors are slightly informative, as is well known in weak lensing [7, 33, 34].

Next, we analyse the 100.000 simulations purposefully with a biased covariance matrix. We left- and right-multiply the KiDS covariance with a diagonal biasing matrix B, whose diagonal elements are drawn from a Gaussian distribution

$$\mathsf{B} = \mathrm{diag}(b_{11}, b_{22}, ..., b_{nn}), \quad \text{with} \quad b_{ii} \sim \mathcal{G}(1, \sigma^2). \tag{3.20}$$

The larger the standard deviation σ, the larger will be the bias in $\mathsf{C}_{\mathrm{B}}^{-1} = \mathsf{B}^\top \mathsf{C}^{-1} \mathsf{B}$. The relative difference between original covariance matrix, and biased covariance matrix then follows from the mean and standard deviations of $\mathsf{B}^\top \mathsf{C}^{-1} \mathsf{B}$. Per matrix element, we have on

average

$$\langle (\mathsf{C}_{\mathrm{B}}^{-1})_{ij} \rangle = (\mathsf{C}^{-1})_{ij} \langle b_{ii} b_{jj} \rangle$$
$$= (\mathsf{C}^{-1})_{ij} \tag{3.21}$$

and using $\langle b_{ii}^2 \rangle = \sigma^2 + 1 \; \forall i$, the variance follows to be

$$\mathrm{Var}[(\mathsf{C}_{\mathrm{B}}^{-1})_{ij}] = \langle [(\mathsf{C}^{-1})_{ij}]^2 \rangle - (\mathsf{C}^{-1})_{ij}^2$$
$$= (\mathsf{C}^{-1})_{ij}^2 [\sigma^4 + 2\sigma^2]. \tag{3.22}$$

According to equation 3.21 the bias vanishes on average, and has according to equation 3.22 a standard deviation of $s = \mathsf{C}_{ij}^{-1} \sqrt{\sigma^4 + 2\sigma^2}$. The relative difference r between biased and correct covariance matrix is then

$$r = \sqrt{\sigma^4 + 2\sigma^2}, \tag{3.23}$$

which is independent of matrix indices ij. The relative differences r can be compared to the literature: for example, the DES reanalysis of the KiDS-450 data [30] implemented 40 percent changes in the covariance matrix elements (see Figure 1 of [30]). Current approximate covariance matrices in weak lensing are therefore uncertain to approximately a degree of $r \approx 0.4$.

We study such example biases in figure 6, for $r = 0.07$, $r = 0.14$ and $r = 0.36$. Analyzing the data with such biased covariance matrices causes the posterior to preferentially peak in the wrong region of parameter space, thereby excluding the true cosmology too often. The left panel of figure 6 illustrates this effect by showing how the best-fitting cosmologies are shifted. To each of these new best-fitting cosmologies belongs a new posterior (not plotted) whose credibility contours are of approximately the same shape as those of the original posterior, only centered on the new best fits. The blue and pink coverage measurements in the right panel of figure 6 indicate how quickly the posterior begins to undercover if the biases of such covariance matrices are not mitigated.

Since the left of figure 6 indicates that for current levels of covariance matrix uncertainty ($r = 0.36$) the best fit scatters over nearly the entire undebiased posterior, we conclude that such uncertainties definitely need to be propagated. We illustrate such a propagation first for the traditional Bayesian marginalization, and then for coverage correction.

Given our bias model with B, the posterior of cosmological parameters when marginalized over B is given by

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \int \mathcal{G}\left(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{x}, \mathsf{B}^T \mathsf{C}^{-1} \mathsf{B}\right) \pi(\mathsf{B}) \mathrm{d}\mathsf{B}, \tag{3.24}$$

where the uncertainty of B is

$$\pi(\mathsf{B}) = \prod_i \mathcal{G}(1, \sigma^2). \tag{3.25}$$

The matrix-variate integration dB is element-wise which becomes quickly numerically prohibitive due to the curse of dimensionality. For the 130-dimensional diagonal B used in equation 3.20, it is still feasible, and we implement it via a Monte-Carlo integration. The resulting posterior, $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$, is depicted in figure 7 in blue dotted contours, and is wider than the original KiDS posterior (solid grey contours) due to the marginalization.

The Bayesian marginalization was here only possible since we knew the model which caused the bias. In a realistic analysis, such a model is not known, and we need to propagate the bias blindly via coverage calibration.
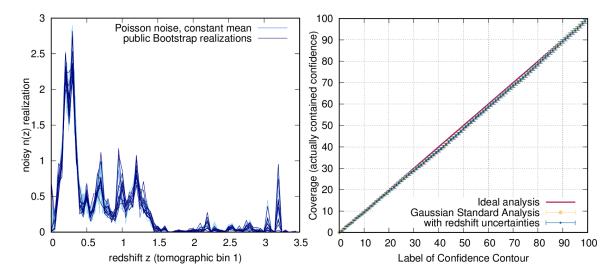
Figure 8: Left: Examples of random realizations of noisy redshift distributions, here for the first tomographic redshift bin from KiDS-450. Right: The coverage is essentially unaffected by redshift noise, illustrating that redshift noise is in current analyses sub-dominant to cosmic variance and shape noise.

We therefore compute the posterior

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}, \mathsf{B}^T\mathsf{C}^{-1}\mathsf{B}), \tag{3.26}$$

of which we know that it must be biased to unknown degree, due to having used the covariance matrix of unknown bias. The measured coverages in figure 6 reveal that for $\sigma = 0.25(r = 0.36)$, the credibility contour which contains 68% of the posterior volume, only covers the true cosmology 42% of the time. In contrast, the contour which contains 92% of the posterior volume, included the true cosmology 68% of the time. The 92% credibility contour of the biased posterior is hence only the 68% credibility contour after debiasing. Figure 7 shows that this coverage calibration complies with the Bayesian marginalization, with the advantage that it required no parametric model.

3.5 Mitigating redshift uncertainties by coverage calibration

In this section, the aim is to propagate redshift uncertainties in a non-parametric manner, for the following reasons.

Estimating the redshift z of a galaxy becomes difficult when only photometric flux measurements are available. Tomographic weak lensing analyses assign galaxies to distributions $n_i(z)$, where i denotes the bin, $i \in [1, r]$. If the redshifts have to be determined photometrically, then the estimated galaxy distributions are uncertain and we write $\hat{n}_i(z)$. There will thus exist a probability distribution

$$\mathcal{P}(\{\hat{n}_i(z)\}|\{n_i(z)\}), \tag{3.27}$$

where the curly braces indicate the set of all tomographic bins.

Propagating uncertainty from the $\hat{n}_i(z)$ through weak lensing analyses is difficult. Bayesian analyses would try to establish the precise functional form of $\mathcal{P}$ in equation 3.27, and then

marginalize over it, resulting in the marginal posterior of cosmological parameters $\boldsymbol{\theta}$

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{\xi}_{\pm}) = \int \mathcal{P}\left(\boldsymbol{\theta}, \boldsymbol{\xi}_{\pm}, \{\hat{n}_i(z)\}, \mathsf{C}\right) \mathcal{P}(\{\hat{n}_i(z)\}|\{n_i(z)\}) \, \mathrm{d}n_1...\mathrm{d}n_r. \qquad (3.28)$$

This integral is numerically extremely costly[5], and has to date not yet been solved. Consequently, the current standard approach is to introduce nuisance parameters instead.

Both KiDS and DES introduce nuisance parameters which shift the centers of redshift bins [8, 35]. It has however been found [35], that the nuisance parameter originally introduced for the intrinsic alignment amplitude also fits to redshift uncertainties. This problem occurs because nuisance parameters *fit*, i.e. they are part of an inverse problem which enables them to compensate for unintended systematics. Note also, that the adopted parametric nuisance model is limited in the sense of not being able to create deformations of $\hat{n}_i(z)$ which leave the central redshifts invariant.

We therefore wish to study the impact of redshift uncertaintites in isolation. Consequently, we replace the nuisance parameters by a forward model of redshift noise. Since the forward model generates redshift noise only, a confusion with intrinsic alignments is excluded. We then use coverage calibration to propagate the redshift uncertainties into the cosmological parameters.

To implement $\mathcal{P}(\{\hat{n}_i(z)\}|\{n_i(z)\})$, we use the published redshift uncertainties from KiDS-450. We use the weighted direct calibration 'DIR' setup of KiDS, which matches spectroscopic galaxy observations and galaxies seen in KiDS. On average, DIR causes approximately 20% uncertainties in each point, but we use the exact errors per point.

We implement four different forward models for redshift uncertainties. The first model generates functions $\hat{n}_i(z)$ whose shape and mean vary. The second model varies the shape only, but keeps the central redshift fixed. This generates uncertainties which cannot be modelled by marginalizing over the mean redshift. For both cases, we use two noise processes: Poisson realizations and the public Bootstrap realizations from KiDS [7].

Examples of the resulting noisy redshift distributions are shown in the left panel of figure 8. For each of these, we compute the theoretical prediction for the KiDS-450 data vector, and analyze it with a Gaussian likelihood, using the public KiDS-covariance matrix. The right panel of Figure 8 reveals that none of the four noise models caused the posteriors to undercover – this means that reported problems with redshifts in KiDS-450 must arise from a bias, or confusion with another systematic effect. Redshift noise in isolation, as here studied, seems to be subdominant to shape noise and cosmic variance, as included in the covariance matrix.

4 Forecasts for dark energy constraints with a Euclid-like survey

As the precision of cosmic surveys improves, the relative impact of formerly negligible biases increases. The upcoming Euclid survey [13], but also its sibling surveys LSST and WFIRST [36], will study the cosmological standard model, and its constituents. The cosmological standard model ΛCDM is based on a cosmological constant Λ and cold dark matter (CDM). In ΛCDM, Λ takes the role of dark energy, and physics beyond the standard model accordingly often introduces additional parameters, w_0 and w_a, for extended dark energy phenomenology [26, 27]. In ΛCDM, these parameters take values $w_0 = -1$, and $w_a = 0$. If the upcoming

---

[5]Already figure 8 required a CPU-time of 26 days in parallel on 10 modern Xenon CPUs (100 times as long as a computation for a single redshift realization).

Table 2: We list which posterior volume contours can be guaranteed to contain at least the probability stated, thereby guaranteeing the absence of biases in constraints on the dark energy parameters w_0 and w_a. The number of simulations is N. The percent probability that shall be guaranteed to be contained in a certain contour is listed in the table header. The body list which posterior volume contour is guaranteed to contain at least this probability, or more. If infinitely many accurate simulations are run, then posterior volume can be guaranteed to directly measure probability (first line). For fewer simulations (remaining lines), biases are possible, such that contours can only be guaranteed to contain a somewhat lower probability than their encased volume would suggest. As an example, for 225 simulations, a bias is either found and mitigated, or if no bias is found, then the 92.5% posterior volume contour can be guaranteed to contain the true parameters at least 90% of the time. Accordingly, if ΛCDM were excluded by the 92.5% contour, then a new standard model might be considered. The last column indicates the percentual decrease of the figure of merit, as a function of number of simulations.

N	$\sqrt{N}$	$\geq 68\%$ probability	$\geq 90\%$ probability	$\geq 95\%$ probability	ΔFoM
∞	∞	68% volume cont.	90.0% volume cont.	95.0% volume cont.	0%
625	25	70% volume cont.	91.5% volume cont.	96% volume cont.	6%
400	20	71% volume cont.	92.0% volume cont.	96% volume cont.	8%
225	15	72% volume cont.	92.5% volume cont.	96.5% volume cont.	10%
100	10	73% volume cont.	93% volume cont.	97% volume cont.	12%
49	7	75% volume cont.	94% volume cont.	97.5% volume cont.	17%
25	5	77% volume cont.	95% volume cont.	98% volume cont.	22%

Euclid analyses exclude this point with high significance, then ΛCDM is ruled out and a new standard model is needed – or a bias occurred.

Due to the complexity of the data analysis, the occurrence of an unrecognized bias is of course possible, but our method is able to tell these biases and new physics apart.

We imagine a Euclid-like survey develops a likelihood, which is as accurate as possible, and which does not rely on simulations. If the likelihood is very accurate, then our method will need to correct only minor outstanding biases, resulting in a minor increase of credibility contours. This likelihood is then to be augmented by few, but very accurate, end-to-end simulations for ΛCDM. We here forecast the number of simulations needed to guarantee that ΛCDM is not discarded due to unrecognized biases.

According to [13], Euclid's prime scientific target is the determination of the dark energy equation of state parameters w_0 and w_a to a precision of

$$\sigma(w_0) = 0.02, \quad \sigma(w_a) = 0.1, \tag{4.1}$$

where σ is the 1-sigma standard deviation. In a Gaussian approximation, the joint confidence contours of w_0 and w_a are elliptical, and the figure of merit (FoM) measures this ellipses area.

For N simulations, our coverage estimator $\hat{C}_\alpha$ has a standard deviation of $\sigma = \sqrt{\alpha(1-\alpha)/N}$. For N simulations, it will thus detect biases which change confidence contours by more than σ. It cannot detect biases which change the coverage by less than σ, and accordingly

$$\hat{C}_\alpha^{\mathrm{low}} = \hat{C}_\alpha - \sigma, \tag{4.2}$$

is a conservative lower estimate of the coverage probability, to be interpreted as the 'most conservative scenario' of mitigating all possible biases which could not yet be ruled out. A
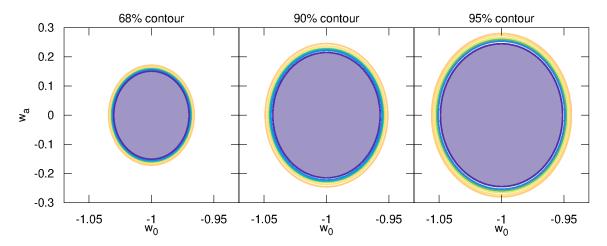
Figure 9: Forecast of conservative credibility contours for the dark energy parameters w_0, w_a. Filled inner contour: Euclid's nominal precision, corresponding to infinitely many simulations and a bias-free inference. Open contours: conservative contours, ranging from 625 (blue) to 25 simulations (orange), spaced as in table 2. Right panel: If the Euclid analysis is augmented by 25 accurate simulations (orange), and no bias needs to be corrected, then its here shown conservative $\geq 95\%$ credibility constraints on w_0 are still approximately 25 times more constraining than current KiDS or DES analyses.

credibility contour which reaches coverage $\hat{C}_\alpha^{\mathrm{low}}$ contains the true parameters *at least* with probability α, and likely more.

In figure 9 we forecast such conservative credibility contours for Euclid. The inner filled contour depicts the 68, 90, and 95 percent credibility contours for Euclid at its nominal precision. These correspond to a bias-free inference with infinitely many simulations. If fewer simulations are available, then the conservative contours increase in size, which is depicted in open contours as a function of N, as given in table 2.

The right panel of figure 9 can be compared to current constraints from DES and KiDS: both of these surveys measure an equation of state w_0, but keeping w_a fixed to its fiducial value. Both surveys currently achieve approximately $-2.0 < w_0 < -0.4$ [7, 8]. Figure 9 therefore indicates that if Euclid's data can be augmented by 25 simulations, then either biases can be detected and mitigated, or if no biases are detected but the shown conservative contours are chosen, then Euclid will still achieve approximately 25 times the precision of KiDS and DES for w_0.

5   Discussion

This paper presented a method to mitigate biases, recognized or unrecognized, even when a Bayesian solution cannot be conducted. Our method takes any existing data analysis pipeline as input, and runs it on simulations and the real data alike. It then measures the *coverage probability* of credibility contours and corrects for it, if found to be off. This produces debiased contours as particle physicists (and many cosmologists) expect them to be: under a repetition of the experiment, the 68 percent confidence contour will contain the true parameters 68 percent of the time, despite the data being analyzed with an imperfect pipeline. Our method can also be understood as a sanity check for any cosmological analysis.

We showed how approximate covariance matrices determine where a likelihood peaks, and that a reduced-χ^2 of order unity does not indicate an unbiased best-fit. To illustrate both points, we forced the original KiDS-450 data set to peak at the Planck best-fitting cosmology, with the exact same χ^2. We then used our method to show how inferences with approximate covariance matrices can be debiased.

We also isolated the impact of uncertain redshifts by using a forward model, since an inverse treatment was found to confuse redshift uncertainties and intrinsic alignments [35]. We found that in isolation, current redshift uncertainties are fully subdominant to shape noise and cosmic variance in current weak lensing analyses. Our study focuses on *uncertaintites* not *biases* in redshifts.

Finally, we illustrated that a pessimistic analysis of Euclid-like data, will very likely constrain the dark energy equation of state parameters by a factor of at least 25 better than current KiDS and DES analyses. This statement assumes that 25 end-to-end simulations of Euclid-like data can be provided alongside an independent likelihood.

This paper, in conjunction with [11], now found repeatedly that *data vector truncation* influences cosmological parameter constraints: the problematic data points always occured at the extreme of angular ranges. This motivates that blinding strategies should be kept for all upcoming surveys.

Our method is applicable to many more examples, and the code is hence public at github.com/elenasellentin/Mitigate_Unrecognized_Biases.

Acknowledgements

References

[1] E. Sellentin and A. F. Heavens, *Parameter inference with estimated covariance matrices*, *MNRAS* **456** (Feb., 2016) L132–L136, [arXiv:1511.05969].

[2] E. Sellentin and A. F. Heavens, *Quantifying lost information due to covariance matrix estimation in parameter inference*, MNRAS **464** (Feb., 2017) 4658–4665, [arXiv:1609.00504].

[3] A. Taylor, B. Joachimi, and T. Kitching, *Putting the precision in precision cosmology: How accurate should your data covariance matrix be?*, MNRAS **432** (July, 2013) 1928–1946, [arXiv:1212.4359].

[4] F. Lacasa and M. Kunz, *Inadequacy of internal covariance estimation for super-sample covariance*, A&A **604** (Aug., 2017) A104, [arXiv:1703.03337].

[5] E. Sellentin, C. Heymans, and J. Harnois-Déraps, *The skewed weak lensing likelihood: why biases arise, despite data and theory being sound.*, MNRAS (Apr., 2018) [arXiv:1712.04923].

[6] C. Hahn, F. Beutler, M. Sinha, A. Berlind, S. Ho, and D. W. Hogg, *Likelihood Non-Gaussianity in Large-Scale Structure Analyses*, *arXiv e-prints* (Mar., 2018) [arXiv:1803.06348].

[7] H. Hildebrandt, M. Viola, C. Heymans, S. Joudaki, K. Kuijken, C. Blake, T. Erben, B. Joachimi, D. Klaes, L. Miller, C. B. Morrison, R. Nakajima, G. Verdoes Kleijn, A. Amon, A. Choi, G. Covone, J. T. A. de Jong, A. Dvornik, I. Fenech Conti, A. Grado, J. Harnois-Déraps, R. Herbonnet, H. Hoekstra, F. Köhlinger, J. McFarland, A. Mead, J. Merten, N. Napolitano, J. A. Peacock, M. Radovich, P. Schneider, P. Simon, E. A. Valentijn,

J. L. van den Busch, E. van Uitert, and L. Van Waerbeke, *KiDS-450: cosmological parameter constraints from tomographic weak gravitational lensing*, MNRAS **465** (Feb., 2017) 1454–1498, [arXiv:1606.05338].

[8] M. A. Troxel, N. MacCrann, J. Zuntz, T. F. Eifler, E. Krause, S. Dodelson, D. Gruen, J. Blazek, O. Friedrich, S. Samuroff, J. Prat, L. F. Secco, C. Davis, J. Weller, and Y. Zhang, *Dark Energy Survey Year 1 Results: Cosmological Constraints from Cosmic Shear*, *ArXiv e-prints* (Aug., 2017) [arXiv:1708.01538].

[9] K. Kuijken, C. Heymans, H. Hildebrandt, R. Nakajima, T. Erben, J. T. A. de Jong, M. Viola, A. Choi, H. Hoekstra, L. Miller, E. van Uitert, A. Amon, C. Blake, and M. Brouwer, *Gravitational lensing analysis of the Kilo-Degree Survey*, MNRAS **454** (Dec., 2015) 3500–3532, [arXiv:1507.00738].

[10] M. Bartelmann and P. Schneider, *Weak gravitational lensing*, Phys. Rep. **340** (Jan., 2001) 291–472, [astro-ph/9912508].

[11] E. Sellentin and A. F. Heavens, *On the insufficiency of arbitrarily precise covariance matrices: non-Gaussian weak-lensing likelihoods*, MNRAS **473** (Jan., 2018) 2355–2363, [arXiv:1707.04488].

[12] J. Harnois-Déraps and L. van Waerbeke, *Simulations of weak gravitational lensing - II. Including finite support effects in cosmic shear covariance matrices*, MNRAS **450** (July, 2015) 2857–2873, [arXiv:1406.0543].

[13] R. Laureijs, J. Amiaux, S. Arduini, J. . Auguères, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, and et al., *Euclid Definition Study Report*, *ArXiv e-prints* (Oct., 2011) [arXiv:1110.3193].

[14] L. Amendola, S. Appleby, A. Avgoustidis, D. Bacon, T. Baker, M. Baldi, N. Bartolo, A. Blanchard, C. Bonvin, S. Borgani, E. Branchini, C. Burrage, S. Camera, C. Carbone, L. Casarini, M. Cropper, C. de Rham, J. P. Dietrich, C. Di Porto, R. Durrer, A. Ealet, P. G. Ferreira, and T. Zlosnik, *Cosmology and fundamental physics with the Euclid satellite*, *Living Reviews in Relativity* **21** (Apr., 2018) 2, [arXiv:1606.00180].

[15] N. Syring and R. Martin, *Calibrating general posterior credible regions*, arXiv:1509.00922.

[16] D. Prangle, M. G. B. Blum, G. Popovic, and S. A. Sisson, *Diagnostic tools of approximate Bayesian computation using the coverage property*, *ArXiv e-prints* (Jan., 2013) [arXiv:1301.3166].

[17] A. F. Heavens and E. Sellentin, *Objective Bayesian analysis of neutrino masses and hierarchy*, J. Cosmology Astropart. Phys. **4** (Apr., 2018) 047, [arXiv:1802.09450].

[18] P. Grünwald and T. van Ommen, *Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it*, *Bayesian Anal.* **12** (12, 2017) 1069–1103.

[19] J. Morrison and N. Simon, *Rank conditional coverage and confidence intervals in high dimensional problems*, *ArXiv e-prints* (Feb., 2017) [arXiv:1702.06986].

[20] M. Goldstein, *Subjective bayesian analysis: Principles and practice*, *Bayesian Anal.* **1** (09, 2006) 403–420.

[21] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan, *Statistical methods for eliciting probability distributions*, *Journal of the American Statistical Association* **100** (2005), no. 470 680–701, [https://doi.org/10.1198/016214505000000105].

[22] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, J. C. Lee, N. J. Nunes, R. Pain, C. R. Pennypacker, R. Quimby, C. Lidman, R. S. Ellis, M. Irwin, R. G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B. J. Boyle, A. V. Filippenko, T. Matheson, A. S. Fruchter, N. Panagia, H. J. M. Newberg, W. J. Couch, and T. S. C.

Project, *Measurements of Ω and Λ from 42 High-Redshift Supernovae*, ApJ **517** (June, 1999) 565–586, [astro-ph/9812133].

[23] A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry, *Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant*, AJ **116** (Sept., 1998) 1009–1038, [astro-ph/9805201].

[24] M. Tegmark, A. N. Taylor, and A. F. Heavens, *Karhunen-Loève Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets?*, ApJ **480** (May, 1997) 22–35, [astro-ph/9603021].

[25] M. Kilbinger, *Cosmology with cosmic shear observations: a review*, Reports on Progress in Physics **78** (July, 2015) 086901, [arXiv:1411.0115].

[26] J. Lesgourgues, *The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview*, ArXiv e-prints (Apr., 2011) [arXiv:1104.2932].

[27] D. Blas, J. Lesgourgues, and T. Tram, *The Cosmic Linear Anisotropy Solving System (CLASS). Part II: Approximation schemes*, J. Cosmology Astropart. Phys. **7** (July, 2011) 034, [arXiv:1104.2933].

[28] R. Takahashi, M. Sato, T. Nishimichi, A. Taruya, and M. Oguri, *Revising the Halofit Model for the Nonlinear Matter Power Spectrum*, ApJ **761** (Dec., 2012) 152, [arXiv:1208.2701].

[29] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al., *Planck 2015 results. XIII. Cosmological parameters*, A&A **594** (Sept., 2016) A13, [arXiv:1502.01589].

[30] M. A. Troxel, E. Krause, C. Chang, T. F. Eifler, O. Friedrich, D. Gruen, N. MacCrann, A. Chen, C. Davis, J. DeRose, S. Dodelson, M. Gatti, B. Hoyle, D. Huterer, M. Jarvis, F. Lacasa, H. V. Peiris, J. Prat, S. Samuroff, C. Sánchez, E. Sheldon, P. Vielzeuf, M. Wang, J. Zuntz, F. B. Abdalla, S. Allam, J. Annis, S. Avila, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, J. De Vicente, H. T. Diehl, P. Doel, A. E. Evrard, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, D. L. Hollowood, K. Honscheid, D. J. James, D. Kirk, K. Kuehn, N. Kuropatkin, T. S. Li, M. Lima, M. March, F. Menanteau, R. Miquel, J. J. Mohr, R. L. C. Ogando, A. A. Plazas, A. Roodman, E. Sanchez, V. Scarpine, R. Schindler, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, D. Thomas, A. R. Walker, and R. H. Wechsler, *Survey geometry and the internal consistency of recent cosmic shear measurements*, ArXiv e-prints (Apr., 2018) [arXiv:1804.10663].

[31] J. Harnois-Deraps, A. Amon, A. Choi, V. Demchenko, C. Heymans, A. Kannawadi, R. Nakajima, E. Sirks, L. van Waerbeke, Y.-C. Cai, B. Giblin, H. Hildebrandt, H. Hoekstra, L. Miller, and T. Troester, *Cosmological Simulations for Combined-Probe Analyses: Covariance and Neighbour-Exclusion Bias*, ArXiv e-prints (May, 2018) [arXiv:1805.04511].

[32] N. MacCrann, J. DeRose, R. H. Wechsler, J. Blazek, E. Gaztanaga, M. Crocce, E. S. Rykoff, M. R. Becker, B. Jain, E. Krause, T. F. Eifler, D. Gruen, J. Zuntz, M. A. Troxel, J. Elvin-Poole, J. Prat, M. Wang, S. Dodelson, A. Kravtsov, P. Fosalba, M. T. Busha, A. E. Evrard, and D. Huterer, *DES Y1 Results: Validating cosmological parameter estimation using simulated Dark Energy Surveys*, ArXiv e-prints (Mar., 2018) [arXiv:1803.09795].

[33] G. Datta and R. Mukerjee, *Probability Matching Priors: Higher Order Asymptotics. Springer, New York.* 2004.

[34] N. MacCrann, J. Zuntz, S. Bridle, B. Jain, and M. R. Becker, *Cosmic discordance: are Planck*

CMB and CFHTLenS weak lensing measurements out of tune?, MNRAS **451** (Aug., 2015) 2877–2888, [arXiv:1408.4742].

[35] E. van Uitert, B. Joachimi, S. Joudaki, A. Amon, C. Heymans, F. Köhlinger, M. Asgari, C. Blake, A. Choi, T. Erben, D. J. Farrow, J. Harnois-Déraps, H. Hildebrandt, H. Hoekstra, T. D. Kitching, D. Klaes, K. Kuijken, J. Merten, L. Miller, R. Nakajima, P. Schneider, E. Valentijn, and M. Viola, *KiDS+GAMA: cosmology constraints from a joint analysis of cosmic shear, galaxy-galaxy lensing, and angular clustering*, MNRAS **476** (June, 2018) 4662–4689, [arXiv:1706.05004].

[36] B. Jain, D. Spergel, R. Bean, A. Connolly, I. Dell'antonio, J. Frieman, E. Gawiser, N. Gehrels, and L. Gladney, *The Whole is Greater than the Sum of the Parts: Optimizing the Joint Science Return from LSST, Euclid and WFIRST*, ArXiv e-prints **1501.07897** (Jan., 2015) [arXiv:1501.07897].