Distribution-free uncertainty quantification for inverse problems: application to weak lensing mass mapping

H. Leterme[1,2], J. Fadili[1], and J.-L. Starck[2,3]

[1] Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France
e-mail: `hubert.leterme@ensicaen.fr`
[2] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191, Gif-sur-Yvette, France
[3] Institutes of Computer Science and Astrophysics, Foundation for Research and Technology Hellas (FORTH), Greece

ABSTRACT

Aims. In inverse problems, distribution-free uncertainty quantification (UQ) aims to obtain error bars in the reconstruction with coverage guarantees that are independent of any prior assumptions about the data distribution. This allows for a better understanding of how intermediate errors propagate through the pipeline. In the context of mass mapping, uncertainties could lead to errors that affects our understanding of the underlying mass distribution, or could propagate to cosmological parameter estimation, thereby impacting the precision and reliability of cosmological models. Current surveys, such as Euclid or Rubin, will provide new weak lensing datasets of very high quality. Accurately quantifying uncertainties in mass maps is therefore critical to fully exploit their scientific potential and to perform reliable cosmological parameter inference.

Methods. In this paper, we extend the conformalized quantile regression (CQR) algorithm, initially proposed for scalar regression, to inverse problems. We compare our approach with another distribution-free approach based on risk-controlling prediction sets (RCPS). Both methods are based on a calibration dataset, and offer finite-sample coverage guarantees that are independent of the data distribution. Furthermore, they are applicable to any mass mapping method, including blackbox predictors. In our experiments, we apply UQ on three mass-mapping method: the Kaiser-Squires inversion, iterative Wiener filtering, and the MCALens algorithm.

Results. Our experiments reveal that RCPS tends to produce overconservative confidence bounds with small calibration sets, whereas CQR is designed to avoid this issue. Although the expected miscoverage rate is guaranteed to stay below a user-prescribed threshold regardless of the mass mapping method, selecting an appropriate reconstruction algorithm remains crucial for obtaining accurate estimates, especially around peak-like structures, which are particularly important for inferring cosmological parameters. Additionally, the choice of mass mapping method influences the size of the error bars.

Key words. cosmology: observations – methods: statistical – gravitational lensing: weak

1. Introduction

Mapping the distribution of matter across the universe is essential for advancing our understanding of its evolution and constraining cosmological parameters. While dark matter constitutes approximately 85% of the observable universe's total mass according to the ΛCDM model, it cannot be directly observed. Instead, its presence can be inferred from gravitational effects, such as the deflection of light rays from distant galaxies. In the weak lensing regime, this phenomenon causes anisotropic stretching of galaxy images, known as shear, which can be used to reconstruct mass maps (for a detailed review on the topic, see Kilbinger 2015). Estimating shear involves detecting statistical anomalies in galaxy ellipticities (Kaiser & Squires 1993). This task is complicated by inherent noise, making weak lensing mass mapping an ill-posed inverse problem without appropriate priors on the matter density field. Furthermore, missing data often occur due to bright objects in the foreground, all the more compromising straightforward solutions. Recent methodologies, including some relying on deep learning, have aimed to enhance mass map reconstructions. However, quantifying uncertainty, particularly with blackbox predictors, remains a significant challenge.

Having accurate uncertainty estimates in inverse problems is crucial for obtaining reliable solutions and contributing to a deeper scientific understanding of our data. In the case of mass mapping, the main motivations to derive uncertainty quantifiers are: i) uncertainties could lead to errors, impacting our analysis of the underlying mass distribution and affecting theories on the nature of dark matter; ii) mass maps are also used for cosmological parameter inference using high-order statistics, where any error could propagate, leading to uncertainties in these parameters; iii) uncertainties may mask physical models of galaxy and cluster growth, affecting our understanding of the processes that determine the evolution of these structures over cosmic time; and iv) uncertainties could introduce bias in the measurements of cluster masses. Therefore, it is essential to not only improve the quality of the mass reconstructions but also to quantify and minimize uncertainties to optimize the potential of future surveys such as Euclid or Rubin.

In this paper, we build on a paradigm called conformal prediction (Vovk et al. 2005; Lei & Wasserman 2014) in the context of weak lensing mass mapping, using simulated data. Specifically, we propose a novel extension of the conformalized quantile regression (CQR) algorithm (Romano et al. 2019) to inverse problems, especially mass mapping. Relying on a calibration set, it is applicable to any prediction method, including blackbox deep learning models, and offers distribution-free, per-pixel coverage guarantees with prescribed confidence levels. For the sake of comparison, we also applied a calibration procedure based on risk-controlling prediction sets (RCPS) (Angelopoulos et al. 2022b) to the problem at hand. We observed that the

latter approach tends to produce overconservative confidence bounds with small calibration sets, unlike CQR. We tested the two calibration methods on three mass mapping techniques: the Kaiser-Squires inversion (Kaiser & Squires 1993), the forward-backward proximal iterative Wiener filtering (Bobin et al. 2012), and the MCALens iterative algorithm (Starck et al. 2021). We evaluated the methods in terms of miscoverage rate and prediction interval size.

This paper is organized as follows. In Sect. 2, we present a state-of-the-art review on the weak lensing inverse mass mapping problem and the currently used quantification methods. Then, Sect. 3 introduces and compares the two distribution-free calibration approaches (CQR and RCPS). In Sect. 4, we describe our experimental settings and present the results, followed by a discussion in Sect. 5. Finally, Sect. 6 concludes the paper.

2. State of the art on the mass mapping inverse problem with uncertainty quantification

Throughout the paper, we use the following notation conventions. Deterministic vectors are denoted with bold lower-case Greek or Latin letters ($\boldsymbol{\kappa}$), while random vectors are represented with bold sans-serif capital letters ($\mathbf{K}$). Deterministic matrices are indicated by standard bold capital letters ($\mathbf{A}$). Furthermore, indexing is done using brackets ($\boldsymbol{\kappa}[k]$, $\mathbf{K}[k]$ or $\mathbf{A}[k, l]$).

The weak lensing mass mapping problem consists in recovering a convergence map $\boldsymbol{\kappa} \in \mathbb{R}^{K^2}$ from an observed, noisy shear map $\boldsymbol{\gamma} \in \mathbb{C}^{K^2}$. Both fields have been discretized over a square grid of size $K \times K$, with $K \in \mathbb{N}$, and are represented as flattened one-dimensional vectors. The relationship between the shear and convergence maps is expressed as follows:

$$\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\kappa} + \boldsymbol{n}, \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{K^2 \times K^2}$ is a known linear operator, referred to as the inverse Kaiser-Squires filter, and $\boldsymbol{n} \in \mathbb{C}^{K^2}$ is the realization of a Gaussian noise $\mathbf{N}$ with zero mean and known diagonal covariance matrix $\boldsymbol{\Sigma_n} \in \mathbb{R}^{K^2 \times K^2}$. A more detailed description of the problem is provided in Appendix A.

2.1. Existing mass mapping methods

The most simple—and widely used—algorithm for performing mass mapping is the Kaiser-Squires (KS) method (Kaiser & Squires 1993), which consists in inverting the linear operator $\mathbf{A}$ in the Fourier space, then applying a Gaussian smoothing to reduce the noise. However, it yields poor results because it does not properly handle noise and missing data. More recent works have proposed to incorporate handcrafted priors in the optimization problem, leading to improved reconstruction. Among those, we can cite an iterative Wiener algorithm based on a Gaussian prior (Bobin et al. 2012), the GLIMPSE2D algorithm based on a sparse prior in the wavelet domain (Lanusse et al. 2016), and the MCALens algorithm (Starck et al. 2021), which considers the combination of a Gaussian component and a sparse component. More details on these methods are given in Appendix B.

The above approaches are essentially model-driven. That is, their design is based on knowledge about the underlying physics, and as such requires handcrafted modeling that may be oversimplifying. Alternatively, data-driven approaches, which take advantage of recent breakthroughs in deep learning, rely on data to learn priors and accurately reconstruct mass maps. Among those, we can cite denoising approaches based on adversarial

networks (Shirasaki et al. 2019, 2021), DeepMass (Jeffrey et al. 2020), which takes as input the Wiener solution and outputs an enhanced point estimate by using a UNet architecture, and DL-Posterior (Remy et al. 2023), which draws samples from the full Bayesian posterior distribution.

2.2. Uncertainty quantification

We consider a point estimate of the convergence map, denoted by $\hat{\boldsymbol{\kappa}} \in \mathbb{R}^{K^2}$, obtained using one of the mass mapping method presented in Sect. 2.1. Uncertainty quantification (UQ) involves estimating lower and upper bounds $\hat{\boldsymbol{\kappa}}^-$ and $\hat{\boldsymbol{\kappa}}^+$, such that the probability of miscoverage remains below a pre-specified threshold $\alpha \in]0, 1[$. The various approaches to achieve this can be categorized into two main types: frequentist (where the ground truth $\boldsymbol{\kappa}$ is a deterministic unknown vector), and Bayesian (where $\boldsymbol{\kappa}$ is the outcome of a random vector $\mathbf{K}$ associated with a given prior distribution). In the following sections, we review both frameworks, discuss their limitations, and introduce the need for calibration.

2.2.1. Frequentist framework

In this framework, the ground truth convergence map $\boldsymbol{\kappa}$ is deterministic. However, due to noise, the observed shear map $\boldsymbol{\gamma}$ is the outcome of a random vector:

$$\boldsymbol{\Gamma} := \mathbf{A}\boldsymbol{\kappa} + \mathbf{N}, \qquad \text{with} \qquad \mathbf{N} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_n}). \tag{2}$$

Consequently, the point estimate $\hat{\boldsymbol{\kappa}}$ and the bounds $\hat{\boldsymbol{\kappa}}^-$ and $\hat{\boldsymbol{\kappa}}^+$, which are computed from $\boldsymbol{\gamma}$, are also outcomes of random vectors, that we respectively denote by $\hat{\mathbf{K}}$, $\hat{\mathbf{K}}^-$, and $\hat{\mathbf{K}}^+$. In this context, we target the following coverage property:

$$\mathbb{P}\left\{\boldsymbol{\kappa}[k] \notin \left[\hat{\mathbf{K}}^-[k], \hat{\mathbf{K}}^+[k]\right]\right\} \leq \alpha, \tag{3}$$

for any pixel $k \in \{1, \ldots, K^2\}$. In this section, we review some methods to achieve this.

Analytical formulation. We assume that the point estimate $\hat{\boldsymbol{\kappa}}$ is obtained with a linear operator:

$$\hat{\boldsymbol{\kappa}} := \mathbf{B}\boldsymbol{\gamma}, \quad \text{thus,} \quad \hat{\mathbf{K}} := \mathbf{B}\boldsymbol{\Gamma}. \tag{4}$$

As shown in Appendix B, this applies to the KS (B.1) and Wiener (B.5) solutions. Plugging (2) into (4) yields

$$\hat{\mathbf{K}} = \mathbf{BA}\boldsymbol{\kappa} + \mathbf{BN}. \tag{5}$$

Then, we can easily show that $\hat{\mathbf{K}}$ follows a multivariate Gaussian distribution:

$$\hat{\mathbf{K}} \sim \mathcal{N}(\mathbf{BA}\boldsymbol{\kappa}, \boldsymbol{\Sigma_{\hat{\kappa}}}), \qquad \text{with} \qquad \boldsymbol{\Sigma_{\hat{\kappa}}} := \mathbf{B}\boldsymbol{\Sigma_n}\mathbf{B}^*. \tag{6}$$

Now, we consider the following hypothesis:

Hypothesis 1. *The estimator $\hat{\mathbf{K}}$ is unbiased:* $\mathbf{BA}\boldsymbol{\kappa} = \boldsymbol{\kappa}$.

We consider the (deterministic) residual vector $\hat{\boldsymbol{r}}$ satisfying

$$\hat{\boldsymbol{r}}[k] := \Phi_k^{-1}(1 - \alpha/2) > 0, \tag{7}$$

where Φ_k denotes the cumulative distribution function (CDF) of a Gaussian distribution with zero mean and variance

$$\sigma_k^2 := \boldsymbol{\Sigma_{\hat{\kappa}}}[k, k]. \tag{8}$$

Then, by setting

$$\hat{\mathbf{K}}^- := \hat{\mathbf{K}} - \hat{\boldsymbol{r}} \quad \text{and} \quad \hat{\mathbf{K}}^+ := \hat{\mathbf{K}} + \hat{\boldsymbol{r}}, \tag{9}$$

we can prove that, under Hyp. 1, the coverage property (3) is satisfied.

On the other hand, GLIMPSE2D and MCALens (B.6) are associated with nonlinear operators on the form

$$\hat{\boldsymbol{\kappa}} := \mathbf{B}(\boldsymbol{\gamma}) \cdot \boldsymbol{\gamma}, \tag{10}$$

where the matrix $\mathbf{B}(\boldsymbol{\gamma})$ is characterized by a set of active coefficients in a wavelet dictionary, which are dependent on the input $\boldsymbol{\gamma}$. Therefore, $\mathbf{B}(\boldsymbol{\gamma})$ is the outcome of a random matrix $\mathbf{B}(\boldsymbol{\Gamma})$, which we assume to be noise-insensitive, essentially depending on the true convergence map $\boldsymbol{\kappa}$:

Hypothesis 2. $\mathbf{B}(\boldsymbol{\Gamma}) = \mathbf{B}(\mathbf{A}\boldsymbol{\kappa})$.

Under Hyp. 2, $\mathbf{B}(\boldsymbol{\Gamma})$ is approximated with a deterministic matrix $\mathbf{B}(\mathbf{A}\boldsymbol{\kappa})$, allowing us to adopt a similar approach as in the linear case (4), using $\mathbf{B} := \mathbf{B}(\mathbf{A}\boldsymbol{\kappa})$.

Practical implementations. Estimating confidence bounds in the above framework only requires to compute the diagonal elements of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\kappa}}}$, as evidenced in (8). However, we need explicit access to all the elements of $\mathbf{B}$, which is infeasible in practice.

In the KS case, by exploiting the spectral properties $\mathbf{B}$, we can show that

$$\mathrm{diag}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\kappa}}}) = \mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{n}}) * |\boldsymbol{b}|^2, \tag{11}$$

where $\boldsymbol{b}$ denotes the first column vector of $\mathbf{B}$ and $*$ denotes the 2D circular convolution product.[1] In practice, the convolution filter $|\boldsymbol{b}|^2$ is fast-decaying, and therefore can be cropped to a much smaller size with negligible impact on the result.

In a more general case, the diagonal elements of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\kappa}}}$ cannot simply be obtained with a 2D convolution. Alternatively, they can be estimated with a Monte-Carlo approach, by noticing that $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\kappa}}}$ is also the covariance matrix of $\mathbf{BN}$. Then, by propagating noise realizations through operator $\mathbf{B}$, and by computing the empirical variance of the outputs at the pixel level, we get an empirical, unbiased estimate of the diagonal elements of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\kappa}}}$. This method is straightforward for linear operators such as KS and Wiener filters, and has also been used by Starck et al. (2021) for MCALens.

2.2.2. Bayesian framework

Bayesian UQ pursue a different objective than (3). In this framework, the ground truth convergence map $\boldsymbol{\kappa}$ is the outcome of a random vector $\mathbf{K}$ with unknown distribution μ^*. In this context, the inverse problem (2) becomes

$$\boldsymbol{\Gamma} := \mathbf{A}\mathbf{K} + \mathbf{N}. \tag{12}$$

We consider a prior distribution μ estimating μ^*, that can be either built from expert knowledge (model-driven approaches), or learned from a training set (data-driven approaches). We assume μ to be associated with a probability density function $f_\mu(\boldsymbol{\kappa})$. Then, given an observation $\boldsymbol{\gamma}$ drawn from $\boldsymbol{\Gamma}$, one can derive a posterior density $f_\mu(\boldsymbol{\kappa}\,|\,\boldsymbol{\gamma})$ satisfying Bayes' rule:

$$f_\mu(\boldsymbol{\kappa}\,|\,\boldsymbol{\gamma}) \propto f_{\boldsymbol{\Gamma}|\mathbf{K}}(\boldsymbol{\gamma}\,|\,\boldsymbol{\kappa})\, f_\mu(\boldsymbol{\kappa}), \tag{13}$$

---

[1] The 2D convolution is performed after having reshaped the vectors to $K \times K$ matrices.

where the likelihood density $f_{\boldsymbol{\Gamma}|\mathbf{K}}(\boldsymbol{\gamma}\,|\,\boldsymbol{\kappa})$ corresponds to a multivariate Gaussian distribution with mean $\mathbf{A}\boldsymbol{\kappa}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{n}}$, according to (12):

$$f_{\boldsymbol{\Gamma}|\mathbf{K}}(\boldsymbol{\gamma}\,|\,\boldsymbol{\kappa}) \propto \exp\left(-\frac{1}{2}\,\|\boldsymbol{\gamma} - \mathbf{A}\boldsymbol{\kappa}\|_{\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}}^2\right). \tag{14}$$

Now, we consider uncertainty bounds $\hat{\boldsymbol{\kappa}}^-$ and $\hat{\boldsymbol{\kappa}}^+$ satisfying

$$\mathbb{P}_\mu\Big\{ \mathbf{K}[k] \notin [\hat{\boldsymbol{\kappa}}^-[k],\, \hat{\boldsymbol{\kappa}}^+[k]] \,\Big|\, \boldsymbol{\Gamma} = \boldsymbol{\gamma} \Big\} \le \alpha, \tag{15}$$

where the miscoverage rate for pixel $k \in \{1, \ldots, K^2 - 1\}$ is obtained by marginalizing the posterior density over all other pixels:

$$\mathbb{P}_\mu\Big\{ \mathbf{K}[k] \notin [a, b] \,\Big|\, \boldsymbol{\Gamma} = \boldsymbol{\gamma} \Big\}$$
$$:= 1 - \int_{\mathbb{R}^{k-1}} \int_a^b \int_{\mathbb{R}^{K^2-k}} f_\mu(\boldsymbol{\kappa}'\,|\,\boldsymbol{\gamma})\, \mathrm{d}\boldsymbol{\kappa}'. \tag{16}$$

As in Sect. 2.2.1, we denote by $\hat{\mathbf{K}}^-$ and $\hat{\mathbf{K}}^+$ the random variables from which $\hat{\boldsymbol{\kappa}}^-$ and $\hat{\boldsymbol{\kappa}}^+$ are drawn, which are dependent on $\boldsymbol{\Gamma}$.

In practice, computing the full posterior is intractable. Instead, approximate error bars can be obtained using two different families of methods.

Full posterior sampling. There exists a broad literature focusing on sampling high-dimensional posterior distributions, using proximal MCMC algorithms with Langevin dynamics (Pereyra 2016; Durmus et al. 2018; Cai et al. 2018a; Pereyra et al. 2020; Laumont et al. 2022; McEwen et al. 2023; Klatzer et al. 2024), or deep generative models based on neural score matching (Remy et al. 2023), as reviewed in Sect. 2.1. From these samples, we can derive a point estimate $\hat{\boldsymbol{\kappa}}$, corresponding to an empirical approximation of the posterior mean:

$$\hat{\boldsymbol{\kappa}} \approx \mathbb{E}_\mu[\mathbf{K}\,|\,\boldsymbol{\Gamma} = \boldsymbol{\gamma}] := \int_{\mathbb{R}^{K^2}} f_\mu(\boldsymbol{\kappa}'\,|\,\boldsymbol{\gamma})\, \boldsymbol{\kappa}'\, \mathrm{d}\boldsymbol{\kappa}', \tag{17}$$

by computing the pixelwise empirical mean, as well as confidence bounds $\hat{\boldsymbol{\kappa}}^-$ and $\hat{\boldsymbol{\kappa}}^+$, by computing the pixelwise $(\alpha/2)$-th and $(1 - \alpha/2)$-th empirical quantiles, respectively.

Fast Bayesian UQ. MCMC sampling methods offer a detailed representation of the posterior distribution, but are known to be computationally expensive. Alternatively, other approaches can estimate pixelwise error bars orders of magnitude faster without the need for high-dimensional sampling.

For instance, assuming an explicit log-concave prior, concentration inequalities (Pereyra 2017) can be used to provide a stable, though somewhat conservative, approximation of the highest-posterior density region from the MAP point estimate. This method has been used to compute Bayesian error bars in the contexts of radio-interferometry (Cai et al. 2018b) and mass mapping (Price et al. 2020). Recently, Liaudat et al. (2023) developed a data-driven model based on similar principles, implementing a learned prior specifically designed to be log-concave.

Alternatively, Jeffrey & Wandelt (2020) proposed a direct estimation of lower-dimensional marginal posterior distributions (for instance, per-pixel estimation) that can quantify uncertainty without relying on high-dimensional MCMC sampling.

2.2.3. Limits of the methods

The methods reviewed above provide an initial guess for the confidence bounds $\hat{\kappa}^-$ and $\hat{\kappa}^+$, but their accuracy strongly depends on the choice of the reconstruction method and / or the prior data distribution. Let us review the weaknesses of both frameworks more specifically.

Frequentist framework. As explained hereafter, Hyp. 1 generally does not hold. That is, applying the mass mapping method on a noise-free shear map $\mathbf{A}\kappa$ does not necessarily accurately recover κ.

Regarding the KS filter, we have $\mathbf{B} := \mathbf{SA}$, where $\mathbf{S}$ denotes a Gaussian low-pass filter. This smoothing operator induces a bias in the solution. In addition, the Wiener and MCALens solutions can be interpreted as maximum a posteriori (MAP) Bayesian estimates, which are purposely biased:

$$\hat{\kappa} \in \text{argmin}_{\kappa'} \left\{ -\log f_\mu(\kappa' \,|\, \boldsymbol{\gamma}) \right\} \tag{18}$$

$$= \text{argmin}_{\kappa'} \left\{ \frac{1}{2} \left\| \boldsymbol{\gamma} - \mathbf{A}\kappa' \right\|_{\Sigma_n^{-1}}^2 - \log f_\mu(\kappa') \right\}, \tag{19}$$

where $f_\mu(\kappa')$ and $f_\mu(\kappa' \,|\, \boldsymbol{\gamma})$ respectively denote the prior and posterior densities, as described in Sect. 2.2.2. The MAP estimator aims to find the most probable solution according to the prior distribution μ, given an observation $\boldsymbol{\gamma}$. Intuitively, if this prior does not align with the true distribution μ^* from which $\mathbf{K}$ is drawn, then the estimator is likely to produce a poor reconstruction $\hat{\kappa}$ of the ground truth κ, even in the absence of noise. For example, it is well known that convergence maps are poorly modeled by Gaussian distributions (Starck et al. 2021). Consequently, the Wiener filter, which assumes a Gaussian prior, tends to blur reconstructed convergence maps around peak-like structures.

Remark 1. Bayesian MAP interpretations of regularized variational problems should be approached with caution. In fact, this is a possible but not the only interpretation. This is for instance the case for sparse regularization, as employed in GLIMPSE2D and MCALens. We refer readers to Starck et al. (2013) for a detailed discussion on this topic.

We identify two additional limitations of the frequentist approach. First, in nonlinear cases such as MCALens, Hyp. 2 is only an approximation. Second, the Monte-Carlo estimation of $\text{diag}(\Sigma_{\hat{\kappa}})$ may introduce sampling errors. To correct these errors without the computational burden of increasing the number of samples, the confidence bounds can be adjusted using bootstrapping. This approach, however, is beyond the scope of this paper. Instead, we focus on distribution-free calibration methods, as described in Sect. 3.

For all these reasons, the coverage property for the frequentist framework, stated in (3), is no longer guaranteed. As evidenced in our experiments (see Table D.1), this approach tends to underestimate the size of the error bars.

Bayesian framework. If the prior distribution μ, from which the confidence bounds $\hat{\kappa}^-$ and $\hat{\kappa}^+$ are estimated, does not align with the true unknown distribution μ^* associated with $\mathbf{K}$, then the empirical miscoverage rate may be inconsistent with (15), leading to under- or over-confident predictions. In particular, the data-driven methods rely on both the quality of the training data and the ability of the model to capture the correct prior.

In Sect. 3, we will introduce two post-processing calibration procedures, respectively based on conformalized quantile regression (CQR) and risk-controlling prediction sets (RCPS), adjusting the confidence bounds $\hat{\kappa}^-$ and $\hat{\kappa}^+$, obtained in either frequentist or Bayesian frameworks. The goal is to get coverage guarantees that do not suffer from the above limitations. Both methods are distribution-free (that is, they do not require any prior assumption on the data distribution), work for any mass mapping method (including blackbox deep-learning models), and provide valid coverage guarantees in finite samples.

3. Distribution-free uncertainty quantification

In Sect. 3.1, we present the CQR algorithm by Romano et al. (2019), in a generalized formulation that we established. Originally designed for scalar regression, we propose in Sect. 3.2 an extension to inverse problems, including mass mapping. Then, in Sect. 3.3, we review the RCPS algorithm by Angelopoulos et al. (2022b). In Sect. 3.4, we present the major differences between the two approaches. Finally, in Sect. 3.5, we explain how the theoretical guarantees provided by CQR and RCPS differ from those targeted by frequentist and Bayesian UQ methods reviewed in Sect. 2.2.

3.1. Conformalized quantile regression: general framework

Conformal prediction, used in the context of classification (Sadinle et al. 2019; Romano et al. 2020; Angelopoulos et al. 2022a) and quantile regression (CQR) (Romano et al. 2019), offers finite-sample coverage guarantees with user-prescribed confidence levels. In this section, we present a generalized version of the CQR algorithm, in which we introduce the concept of calibration function.

Let (X, Y) denote a pair of random variables taking values in $\mathcal{I} \times \mathbb{R}$, where X is a scalar or multivariate variable of observables, and Y is the response variable. We now consider a prediction interval $\hat{C} : \mathcal{I} \to 2^{\mathbb{R}}$. It satisfies, for any $x \in \mathcal{I}$,

$$\hat{C}(x) := \left[\hat{f}(x) - \hat{r}(x), \, \hat{f}(x) + \hat{r}(x) \right], \tag{20}$$

for some prediction functions $\hat{f} : \mathcal{I} \to \mathbb{R}$ (producing point estimates), and $\hat{r} : \mathcal{I} \to \mathbb{R}_+$ (producing residuals). In practice, the bounds can be obtained via quantile regression. In this framework, the predictors $\hat{f}^- := \hat{f} - \hat{r}$ and $\hat{f}^+ := \hat{f} + \hat{r}$ are designed to approximate the $(\alpha/2)$-th and $(1 - \alpha/2)$-th quantiles of Y, for a given $\alpha \in \,]0, 1[$. However, the following risk of miscoverage:

$$\mathbb{P}\{\mathsf{Y} \notin \hat{C}(\mathsf{X})\} \leq \alpha \tag{21}$$

is only guaranteed asymptotically, for some specific models, and under regularity conditions (Takeuchi et al. 2006; Steinwart & Christmann 2011).

Now, given a calibration parameter $\lambda \in [a, b] \subseteq \overline{\mathbb{R}}$ (with a and b possibly infinite), we introduce a calibrated prediction interval as follows:

$$\hat{C}_\lambda(x) := \left[\hat{f}(x) - g_\lambda(\hat{r}(x)), \, \hat{f}(x) + g_\lambda(\hat{r}(x)) \right], \tag{22}$$

where

$$g_\lambda : \mathbb{R}_+ \to \mathbb{R}_+ \tag{23}$$

denotes a family of non-decreasing calibration functions parameterized by λ, such that $g_{\lambda_0} = \text{Id}$ for some specific value λ_0 (no

calibration), and producing larger prediction intervals with increasing values of λ:

$$\lambda \leq \lambda' \implies \forall r \in \mathbb{R}_+, \; g_\lambda(r) \leq g_{\lambda'}(r). \tag{24}$$

Therefore, g_λ shrinks the initial prediction interval if $\lambda < \lambda_0$, and expands it if $\lambda > \lambda_0$. Moreover, we assume that $\lambda \mapsto g_\lambda(r)$ is continuous. Examples of such calibration functions are provided in Fig. 4. For instance, in the original paper by Romano et al. (2019), $g_\lambda(r) = \max(r + \lambda, 0)$ for any $\lambda \in \mathbb{R}$ (see Fig. 4a), and therefore,

$$\hat{C}_\lambda(x) := \left[\hat{f}^-(x) - \lambda, \; \hat{f}^+(x) + \lambda \right]. \tag{25}$$

The goal is then to find the smallest value of λ such that (21) is guaranteed with $\hat{C} := \hat{C}_\lambda$. For this, we consider a calibration set $(x_i, y_i)_{i=1}^n$ drawn from $n \in \mathbb{N}$ pairs of random variable $(X_i, Y_i)_{i=1}^n$. We then compute the following conformity scores, for any $i \in \{1, \ldots, n\}$ (the smaller the better):

$$\lambda_i := \min \left\{ \lambda \in [a, b] : y_i \in \hat{C}_\lambda(x_i) \right\} \tag{26}$$

$$= \min \left\{ \lambda \in [a, b] : g_\lambda(\hat{r}(x_i)) \geq \left| \hat{f}(x_i) - y_i \right| \right\}, \tag{27}$$

and the conformity score is set to a (resp. b) if the condition is always (resp. never) met. The existence of $\lambda_i \in [a, b]$ is guaranteed by the continuity of $\lambda \mapsto g_\lambda(r)$. In the case where $g_\lambda(r) := \max(r + \lambda, 0)$ as in the original paper, we have

$$\lambda_i = \left| \hat{f}(x_i) - y_i \right| - \hat{r}(x_i). \tag{28}$$

Then, we compute the $(1 - \alpha)(1 + 1/n)$-th empirical quantile of $(\lambda_i)_{i=1}^n$, denoted by $\lambda^{(\alpha)}$. The random variables from which λ_i and $\lambda^{(\alpha)}$ are drawn are respectively denoted by Λ_i and $\Lambda^{(\alpha)}$. The latter depends on the calibration set $(X_i, Y_i)_{i=1}^n$, but is independent of X and Y. The term $(1 + 1/n)$, which accounts for finite-sample correction, imposes a lower bound for the target error rate: $\alpha \geq 1/(n + 1)$. To ease the flow of reading, we write $X_{n+1} := X$ and $Y_{n+1} := Y$.

The following proposition, for which a proof is provided in Appendix C, is a generalization of Theorem 1 by Romano et al. (2019).

Proposition 1. *If $(X_i, Y_i)_{i=1}^{n+1}$ are drawn exchangeably from an arbitrary joint distribution, and if the conformity scores $(\Lambda_i)_{i=1}^{n+1}$ are almost surely distinct, then,*

$$\alpha - \frac{1}{n + 1} \leq \mathbb{P} \left\{ Y \notin \hat{C}_{\Lambda^{(\alpha)}}(X) \right\} \leq \alpha. \tag{29}$$

As an example, i.i.d. random variables are exchangeable. More generally, exchangeability implies identical distribution but not necessarily independence. The almost-surely-distinct condition on $(\Lambda_i)_{i=1}^{n+1}$ implies that each conformity score is almost surely distinct from $\{a, b\}$, indicating that the predictions can be calibrated. The result stated in (29) includes a lower bound for the probability of miscoverage; therefore, CQR avoids overconservative prediction intervals whenever n is large enough.

3.2. Conformalized mass mapping

The CQR algorithm was initially designed for scalar regression. In the context of mass mapping, where both inputs and outputs are multidimensional, we propose applying CQR to each output pixel individually. This idea was also exploited in a very recent paper by Narnhofer et al. (2024), though it was limited to

Bayesian error quantification. In contrast, our method is a more straightforward extension of the CQR algorithm to inverse problems, which does not restrict to the Bayesian framework. To the best of our knowledge, this direction has not been explored before.

Similarly to Sect. 2.2, we denote by $\hat{\kappa}^-$ and $\hat{\kappa}^+$ the initial lower- and upper-confidence bounds, obtained using any mass mapping method and any UQ approach. The random vectors from which they are drawn are written $\hat{\mathbf{K}}^-$, and $\hat{\mathbf{K}}^+$. We also denote by

$$\hat{\kappa} := \frac{\hat{\kappa}^- + \hat{\kappa}^+}{2} \qquad \text{and} \qquad \hat{r} := \frac{\hat{\kappa}^+ - \hat{\kappa}^-}{2} \tag{30}$$

the corresponding point estimate and residual.[2] We consider a calibration set $(\gamma_i, \kappa_i)_{i=1}^n$, drawn from i.i.d. random pairs of shear and convergence maps $(\Gamma_i, \mathbf{K}_i)_{i=1}^n$ following the same distribution as $(\Gamma, \mathbf{K})$. We will explain in Sect. 4.1 how to get such a calibration set. Then, for each pixel $k \in \{1, \ldots, K^2\}$, we apply CQR such as described in Sect. 3.1 with $\mathbb{C}^{K^2}$ as input space $\mathcal{I}$, and

$$x := \gamma, \quad y := \kappa[k], \quad (x_i, y_i)_{i=1}^n := (\gamma_i, \kappa_i[k])_{i=1}^n. \tag{31}$$

The conformity scores are encoded into n score vectors $\lambda_i \in \mathbb{R}^{K^2}$. Similarly, the resulting calibration parameters are represented by a calibration vector $\lambda^{(\alpha)} \in \mathbb{R}^{K^2}$. We denote by $\hat{\kappa}_{\text{cqr}}^-$ and $\hat{\kappa}_{\text{cqr}}^+$ the lower- and upper-bounds obtained by calibrating the initial bounds $\hat{\kappa}^-$ and $\hat{\kappa}^+$ pixelwise:

$$\hat{\kappa}_{\text{cqr}}^- := \hat{\kappa} - g_{\lambda^{(\alpha)}}(\hat{r}) \qquad \text{and} \qquad \hat{\kappa}_{\text{cqr}}^+ := \hat{\kappa} + g_{\lambda^{(\alpha)}}(\hat{r}), \tag{32}$$

where we have defined, for any calibration vector $\lambda \in \mathbb{R}^{K^2}$, any residual $r \in \mathbb{R}^{K^2}$, and any pixel $k \in \{1, \ldots, K^2\}$,

$$g_\lambda(r)[k] := g_{\lambda[k]}(r[k]), \tag{33}$$

for a given scalar calibration function g_λ, as introduced in (23).

Now, we denote by Λ_i, $\Lambda^{(\alpha)}$, $\hat{\mathbf{K}}_{\text{cqr}}^-$, and $\hat{\mathbf{K}}_{\text{cqr}}^+$ the random vectors from which λ_i, $\lambda^{(\alpha)}$, $\hat{\kappa}_{\text{cqr}}^-$, and $\hat{\kappa}_{\text{cqr}}^+$ are drawn, respectively. We assume that the conformity scores $(\Lambda_i[k])_{i=1}^{n+1}$ are almost surely distinct, for each pixel k. Then, by averaging over all pixels, (29) yields the following result:

$$\alpha - \frac{1}{n + 1} \leq \mathbb{E} \left[L \left(\mathbf{K}, \hat{\mathbf{K}}_{\text{cqr}}^-, \hat{\mathbf{K}}_{\text{cqr}}^+ \right) \right] \leq \alpha, \tag{34}$$

where we have defined the imagewise miscoverage rate:

$$L(\kappa, \hat{\kappa}^-, \hat{\kappa}^+) := \frac{\text{card} \left\{ k \in \mathcal{M} : \kappa[k] \notin [\hat{\kappa}^-[k], \hat{\kappa}^+[k]] \right\}}{\text{card} \, \mathcal{M}}, \tag{35}$$

where $\mathcal{M} \subset \{1, \ldots, K^2\}$ denotes a set of active pixels. More details on masked data are provided in Appendix A.4.

---

[2] This point estimate may differ from the one obtained in (17) within the Bayesian framework, because the CQR algorithm presented in this paper requires equally sized lower and upper residuals. Although Romano et al. (2019) proposed an asymmetric version of their algorithm, it tends to increase the length of prediction intervals.

3.3. Risk-controlling prediction sets

In this section, we present an alternative calibration approach based on risk-controlling prediction sets (RCPS). Initially developed by Bates et al. (2021) in the context of classification, it was later extended to inverse problems (Angelopoulos et al. 2022b) and diffusion models (Horwitz & Hoshen 2022; Teneggi et al. 2023). Relying on Hoeffding's inequality (Boucheron et al. 2004), this approach also provides distribution-free, finite-sample coverage guarantees with user-prescribed confidence levels. In contrast to CQR, RCPS aims at controlling the risk of statistical anomalies in the calibration set. However, it does not prevent overconservative confidence intervals.

We consider a family of calibration functions $g_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ as introduced in Sect. 3.1. Then, for a given parameter $\lambda \in \mathbb{R}$, we consider the adjusted bounds $\hat{\kappa}_\lambda^-$ and $\hat{\kappa}_\lambda^+$ satisfying

$$\hat{\kappa}_\lambda^- := \hat{\kappa} - g_\lambda(\hat{r}) \qquad \text{and} \qquad \hat{\kappa}_\lambda^+ := \hat{\kappa} + g_\lambda(\hat{r}). \tag{36}$$

Considering the corresponding random vectors $\hat{\mathbf{K}}_\lambda^-$ and $\hat{\mathbf{K}}_\lambda^+$, we define the risk R_λ as the expected miscoverage rate:

$$R_\lambda := \mathbb{E}\left[L\left(\mathbf{K}, \hat{\mathbf{K}}_\lambda^-, \hat{\mathbf{K}}_\lambda^+ \right) \right], \tag{37}$$

where $L(\kappa, \hat{\kappa}^-, \hat{\kappa}^+)$ have been defined in (35). We seek the smallest value λ satisfying $R_\lambda \leq \alpha$ with some confidence guarantees. For this purpose, we consider a calibration set $(\gamma_i, \kappa_i)_{i=1}^n$ as introduced in Sect. 3.2. For a given error level $\delta \in]0, 1[$, we compute Hoeffding's upper-confidence bound of R_λ, defined by

$$R_{\delta,\lambda}^+\left((\gamma_i, \kappa_i)_{i=1}^n\right) := \frac{1}{n}\sum_{i=1}^n L\left(\kappa_i, \hat{\kappa}_{i,\lambda}^-, \hat{\kappa}_{i,\lambda}^+\right) + \sqrt{\frac{-\log \delta}{2n}}, \tag{38}$$

where $\hat{\kappa}_{i,\lambda}^-$ and $\hat{\kappa}_{i,\lambda}^+$ satisfy (36) with $\hat{\kappa} := \hat{\kappa}_i$ and $\hat{r} := \hat{r}_i$. These estimates are obtained from γ_i with the same method used to compute $\hat{\kappa}$ and $\hat{r}$. The RCPS approach is based on the following property, derived from Hoeffding's inequality:

$$\mathbb{P}\left\{ R_{\delta,\lambda}^+\left((\mathbf{\Gamma}_i, \mathbf{K}_i)_{i=1}^n\right) < R_\lambda \right\} \leq \delta. \tag{39}$$

It controls the risk of underestimating the expected miscoverage rate by computing an empirical estimate on the calibration set.

We consider the following calibration parameter, computed on the calibration set:

$$\lambda^{(\alpha,\delta)} := \inf\left\{ \lambda \in \mathbb{R} : R_{\delta,\lambda}^+\left((\gamma_i, \kappa_i)_{i=1}^n\right) < \alpha \right\}, \tag{40}$$

and the corresponding random vector $\Lambda^{(\alpha,\delta)}$. By construction, the functions $\lambda \mapsto R_\lambda$ and $\lambda \mapsto R_{\delta,\lambda}^+((\gamma_i, \kappa_i)_{i=1}^n)$ are monotone and non-increasing. Furthermore, we assume that the set $\{\lambda \in \mathbb{R} : R_\lambda \leq \alpha\}$ is non-empty. Then, the following result has been proven by Bates et al. (2021), using (39):

$$\mathbb{P}\left\{ R_{\Lambda^{(\alpha,\delta)}} > \alpha \right\} \leq \delta. \tag{41}$$

Now, we consider the calibrated bounds with respect to $\lambda^{(\alpha,\delta)}$:

$$\hat{\kappa}_{\mathrm{rc}}^- := \hat{\kappa}_{\lambda^{(\alpha,\delta)}}^- \qquad \text{and} \qquad \hat{\kappa}_{\mathrm{rc}}^+ := \hat{\kappa}_{\lambda^{(\alpha,\delta)}}^+, \tag{42}$$

where $\hat{\kappa}_{\lambda^{(\alpha,\delta)}}^-$ and $\hat{\kappa}_{\lambda^{(\alpha,\delta)}}^+$ satisfy (36). The corresponding random vectors are denoted by $\hat{\mathbf{K}}_{\mathrm{rc}}^-$ and $\hat{\mathbf{K}}_{\mathrm{rc}}^+$, respectively. The random variable $\Lambda^{(\alpha,\delta)}$ only depends on the calibration set $(\mathbf{\Gamma}_i, \mathbf{K}_i)_{i=1}^n$, and therefore is independent of the miscoverage rate $L(\mathbf{K}, \hat{\mathbf{K}}_\lambda^-, \hat{\mathbf{K}}_\lambda^+)$,

for any $\lambda \in \mathbb{R}$. Consequently, the risk R_λ defined in (37) can be expressed as follows:

$$R_\lambda = \mathbb{E}\left[L\left(\mathbf{K}, \hat{\mathbf{K}}_{\Lambda^{(\alpha,\delta)}}^-, \hat{\mathbf{K}}_{\Lambda^{(\alpha,\delta)}}^+ \right) \mid \Lambda^{(\alpha,\delta)} = \lambda \right] \tag{43}$$

$$= \mathbb{E}\left[L\left(\mathbf{K}, \hat{\mathbf{K}}_{\mathrm{rc}}^-, \hat{\mathbf{K}}_{\mathrm{rc}}^+ \right) \mid \Lambda^{(\alpha,\delta)} = \lambda \right]. \tag{44}$$

Finally, plugging (44) into (41) yields the following coverage guarantee for RCPS:

$$\mathbb{P}\left\{ \mathbb{E}\left[L\left(\mathbf{K}, \hat{\mathbf{K}}_{\mathrm{rc}}^-, \hat{\mathbf{K}}_{\mathrm{rc}}^+ \right) \mid \Lambda^{(\alpha,\delta)} \right] > \alpha \right\} \leq \delta. \tag{45}$$

3.4. Differences between CQR and RCPS

The theoretical coverage guarantees provided by the CQR and RCPS algorithms, respectively, (34) and (45), present some fundamental differences. First, only CQR prevents overconservative solutions, due to the lower bound in (34). Moreover, in both cases, there exists a nonzero probability of accidentally selecting an out-of-distribution calibration set, leading to miscalibration and potentially large error rates. This possibility does not contradict the theoretical guarantees of either CQR or RCPS, because both approaches treat the calibration set as a set of random vectors. However, they handle this situation differently: the expected miscoverage rate is evaluated over the full distribution of calibration sets in the case of CQR, and conditionally on the calibration parameter $\Lambda^{(\alpha,\delta)}$—which depends on the calibration set $(\mathbf{\Gamma}_i, \mathbf{K}_i)_{i=1}^n$—in the case of RCPS. Consequently, unlike CQR, the risk of miscalibration due to selecting a too-small calibration parameter is explicit in (45), and is controlled by the parameter δ. Therefore, RCPS provides a higher degree of control over uncertainties. However, as confirmed by our experiments, it tends to produce more conservative prediction bounds than CQR, even for large values of δ.

Finally, it is worth noticing that mass mapping with CQR, described in Sect. 3.2, requires K^2 calibration parameters gathered in the vector $\lambda^{(\alpha)}$, which can be computed efficiently using linear algebra libraries such as NumPy and SciPy in Python. In contrast, the approach based on RCPS, described in Sect. 3.3, only requires one scalar calibration parameter $\lambda^{(\alpha,\delta)}$.

3.5. Differences with uncalibrated coverage guarantees

By averaging over the set $\mathcal{M}$ of active pixels, the target coverage guarantees before calibration—(3) for the frequentist framework and (15) for the Bayesian framework—become, respectively,

$$\mathbb{E}\left[L\left(\kappa, \hat{\mathbf{K}}^-, \hat{\mathbf{K}}^+ \right) \right] \leq \alpha \tag{46}$$

and

$$\mathbb{E}_\mu\left[L\left(\mathbf{K}, \hat{\mathbf{K}}^-, \hat{\mathbf{K}}^+ \right) \mid \mathbf{\Gamma} = \gamma \right] \leq \alpha, \tag{47}$$

where the conditional expected value $\mathbb{E}_\mu[\cdot \mid \mathbf{\Gamma} = \gamma]$ is defined with respect to the posterior density $f_\mu(\kappa \mid \gamma)$ introduced in (13). In either (34) and (45), the expected miscoverage rate is evaluated over the full distribution of convergence maps, in contrast with (46) or (47), where it is taken conditionally on a single ground truth κ (frequentist case) or observation γ (Bayesian case), respectively. Consequently, the calibration procedures do not prevent off-target miscoverage rates for specific images.
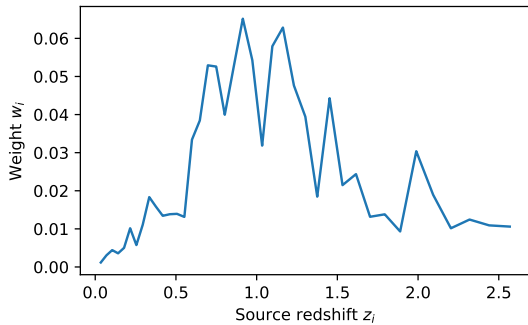
Fig. 1: Distribution of redshifts used to combine simulated convergence maps from the κTNG dataset.

Fig. 2: Number of galaxies per pixel using the S10 weak lensing shear catalog (Schrabback et al. 2010). All redshifts have been considered. The white borders delimitates the survey boundaries. Even within boundaries, some data are missing due to survey measurement masks.

4. Experiments

4.1. Experimental settings

4.1.1. Mass mapping methods

In this paper, we tested CQR and RCPS on three mass mapping methods introduced in Sect. 2.1—the KS inversion, the iterative Wiener algorithm, and MCALens. The uncalibrated uncertainty bounds $\hat{\kappa}^-$ and $\hat{\kappa}^+$ were computed using the frequentist framework presented in Sect. 2.2.1. Specifically, for the KS method, the bounds were obtained analytically using (11). In contrast, for the Wiener and MCALens methods, the bounds were computed using the Monte-Carlo method, by propagating 25 noise realizations through the pipeline for each input shear map. The Python implementation of these algorithms is available on the CosmoStat GitHub repository.[3]

4.1.2. Simulated convergence maps

In order to evaluate and compare the proposed methods, we used a calibration set $(\gamma_i, \kappa_i)_{i=1}^n$ of size $n = 100$ and a test set $(\gamma_i, \kappa_i)_{i=n+1}^m$ of size $m - n = 125$. The ground truths κ_i were obtained using κTNG cosmological hydrodynamic simulations (Osato et al. 2021), which include realizations of $5 \times 5 \deg^2$ convergence maps for various source redshifts at a 0.29 arcmin per pixel resolution, assuming a flat ΛCDM Universe.

For our experiments, we considered linear combinations of convergence maps according to a predefined redshift distribution, that we cropped to 306×306 pixels, with no overlap between them. This corresponds to an opening angle of 1.49 deg. Formally, each convergence map κ_i can be expressed as follows:

$$\kappa_i = \sum_{j=1}^{n_z} w_j \kappa_{ij}, \tag{48}$$

where $(\kappa_{ij})_{j=1}^{n_z}$ denotes a set of convergence map realizations at $n_z = 40$ source redshifts ranging from 3.45×10^{-2} to 2.57, and $w_j \in [0, 1]$ denotes the weight assigned to the j-th source redshift z_j. The weights $(w_j)_{j=1}^{n_z}$, which sum to 1, were chosen to match the shape catalog used to compute the noisy shear maps γ_i (see Sect. 4.1.3). The weight distribution is displayed in Fig. 1.

4.1.3. Noisy shear maps

In accordance with (1), to generate noisy shear maps γ_i from κ_i, we applied the inverse Kaiser-Squires filter $\mathbf{A}$ and added a zero-
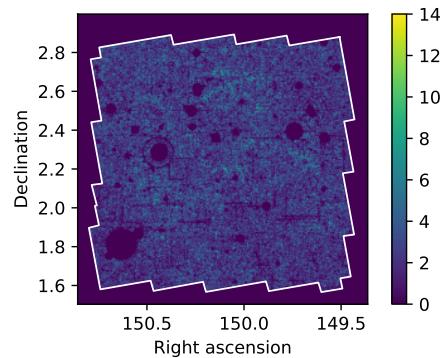
---

mean Gaussian noise $\boldsymbol{n}_i$ with a diagonal covariance matrix $\boldsymbol{\Sigma_n}$ shared among all calibration and test input images. Its diagonal values were obtained by binning the weak lensing shear catalog created by Schrabback et al. (2010) (referred to as "S10", following Remy et al. 2023) at the κTNG resolution (0.29 arcmin per pixel), and then applying (A.14). The intrinsic standard deviation $\sigma_{\rm ell}$ was estimated from the measured ellipticities, and set to 0.28. This is a reasonable estimation, assuming that the shear dispersion is negligible with respect to the dispersion of intrinsic ellipticities. The corresponding number of galaxies per pixel is represented in Fig. 2.

The S10 catalog was based on observations conducted by the NASA/ESA Hubble Space Telescope targeting the COS-MOS field (Scoville et al. 2007), as well as photometric redshift measurements from Mobasher et al. (2007). It contains an average of 32 galaxies per square arcminute over a wide range of redshifts, which is consistent with what is expected from forthcoming surveys like Euclid. This number excludes the galaxies with redshifts below 3.45×10^{-2} and above 2.57, for the sake of consistency with the κTNG dataset. It also ignores the inactive pixels $k \notin \mathcal{M}$, without any observed galaxy. In these areas, we have simply set the shear map values to 0, as explained in Appendix A.4. Consequently, we have omitted these regions from our statistical analyses.

An example of simulated convergence map κ_i and corresponding noisy shear map γ_i is provided in Fig. 3.

4.1.4. Reconstruction parameters

The Wiener estimate $\hat{\kappa}_{\rm wien}$ as well as the Gaussian component $\hat{\kappa}_{\rm G}$ of MCALens have been computed with a power spectrum (diagonal elements of $\mathbf{P}_\kappa$) empirically estimated from a dataset of 180 simulated convergence maps distinct from the calibration and test sets. Moreover, the sparse component $\hat{\kappa}_{\rm S}$ of MCALens was estimated using a starlet dictionary (Starck et al. 2007), which is well-suited for isotropic objects. The detection threshold for selecting the set of active wavelet coefficients was set to 4σ. This threshold was selected based on visual evaluation, as it achieved a good balance between effectively detecting peaks and minimizing false positives. For both iterative methods, the number of iterations was set to the default value of 12. This choice was motivated by seeking a tradeoff between reconstruction accuracy and computational cost. Finally, the Kaiser-Squires maps were
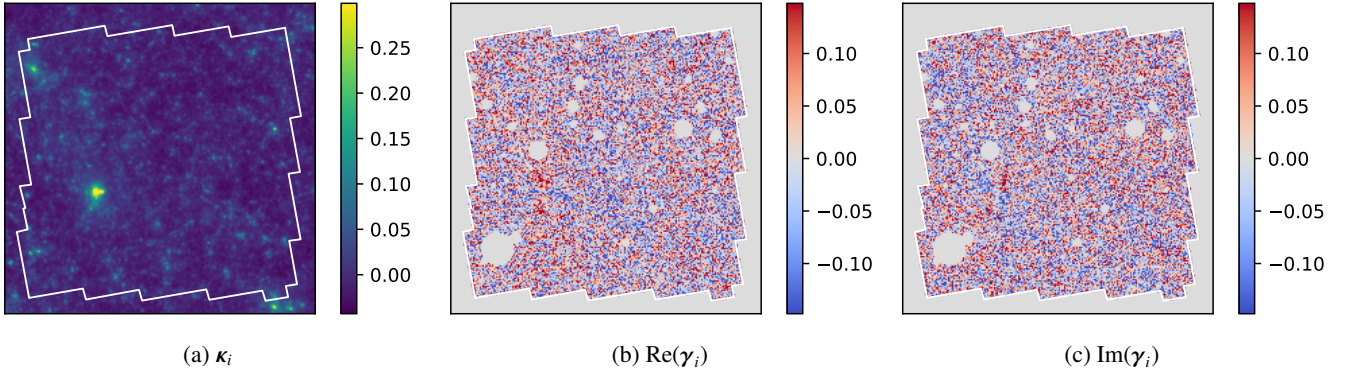
(a) $\boldsymbol{\kappa}_i$ (b) $\mathrm{Re}(\boldsymbol{\gamma}_i)$ (c) $\mathrm{Im}(\boldsymbol{\gamma}_i)$

Fig. 3: Example of simulated convergence map (Fig. 3a), and the corresponding noisy shear map (Figs. 3b and 3c).

smoothed with a Gaussian filter **S** with a full width at half maximum (FWHM) of 2.4 arcmin, following Starck et al. (2021). To test the effect of the low-pass filter on reconstruction accuracy and error bar size, we also used a filter with half the original FWHM. These are referred to as strong and weak smoothing, respectively.

4.1.5. UQ parameters

We selected a target confidence level at 2σ, corresponding to $\alpha \approx 0.046$. We notice that higher confidence levels require larger calibration sets, due to the finite-sample correction: the CQR algorithm selects the $(1 - \alpha)(1 + 1/n)$-th empirical quantile of the conformity scores, which, by definition, must remain below 100%. Consequently, a 2σ-confidence requires at least $n = 21$ calibration samples, versus $n = 370$ at 3σ and $n = 15\,787$ at 4σ. Moreover, regarding the approach based on RCPS, we tested three values for the error level: $\delta = 0.05, 0.2,$ and 0.5.

We implemented both approaches with several families of calibration functions, which is a novel aspect of this work:

$$g_\lambda : r \mapsto \max(r + \lambda, 0) \tag{49}$$

as used by Romano et al. (2019), and

$$g_\lambda : r \mapsto \lambda r \tag{50}$$

as used by Angelopoulos et al. (2022b). The former increases (or decreases) the size of the confidence intervals by the same value λ regardless of the initial size $2r$, whereas the latter adjusts the size proportionally to its initial value. This choice may have an influence on the average size of the calibrated confidence intervals. To push the analysis further, we also tested calibration functions with intermediate behaviors for the CQR approach, in the form

$$g_\lambda : r \mapsto r + b F_{\chi^2(k)}(r/a)(\lambda - 1), \tag{51}$$

where $F_{\chi^2(k)}$ denotes the cumulative distribution function of a chi-squared distribution with k degrees of freedom, and a and b denote positive real numbers. In practice, we tested values of a ranging from 0.004 to 0.012, and adjusted b to the maximum value such that g_λ remains non-decreasing for any $\lambda \geq 0$.

In the rest of the paper, the functions defined in (49), (50) and (51), for which a visual representation is provided in Fig. 4, are referred to as "additive", "multiplicative", and "chi-squared" calibration functions, respectively.

4.2. Results

4.2.1. Visualization of reconstructed convergence maps

A visual example of reconstructed convergence map, together with its prediction bounds before and after calibration, is provided in Fig. D.1 (Appendix D). A focus on the main peak-like structure is displayed in Fig. 8 for the Wiener and MCALens estimates, after calibration with CQR.

We can observe that, for the KS and Wiener methods, the high-density region (bright spot in the convergence map) falls outside the confidence bounds, even after calibration (it is actually under-estimated). This is not in contradiction with the theoretical guarantees stated in (29) and (45), because the expected miscoverage rate, computed across the whole set $\mathcal{M}$ of active pixels, remains below the target α, as explained in Sect. 4.2.3. However, these guarantees do not tell anything about the miscoverage rate of the higher-density regions specifically. A more detailed discussion on this topic is provided in Sect. 5.3.

In contrast, MCALens correctly predicts the high-density region. This observation is consistent with the fact that this algorithm was designed to accurately reconstruct the sparse component of the density field.

4.2.2. Reconstruction accuracy

In order to reproduce previously-established results, we measured, for each $i \in \{n + 1, \dots, m\}$ in the test set, the root mean square error (RMSE) between the ground truth $\boldsymbol{\kappa}_i$ and reconstructed convergence maps $\hat{\boldsymbol{\kappa}}_{\mathrm{ks}\,i}$, $\hat{\boldsymbol{\kappa}}_{\mathrm{wien}\,i}$ and $\hat{\boldsymbol{\kappa}}_{\mathrm{mcal}\,i}$, corresponding to the KS, Wiener and MCALens solutions, respectively. These metrics were computed on the set $\mathcal{M}$ of active pixels, that is, within the survey boundaries, and for pixels with nonzero galaxies. We also measured the RMSE for high-density regions only, which are of greater importance when inferring cosmological parameters. More precisely, we only considered the pixels k such that $|\boldsymbol{\kappa}_i[k]| \geq 4.8 \times 10^{-2}$, which corresponds to a signal-to-noise (S/N) ratio above 0.25 (2.8% of the total number of pixels). The results are displayed in Table 1.

These results confirm that the Wiener and MCALens solutions achieve higher reconstruction accuracy compared to the KS solution, although the latter can be improved by increasing the standard deviation of the smoothing filter (see "KS1" versus "KS2"). Additionally, MCALens outperforms Wiener by 1.6%. This difference significantly increases when focusing on high-density regions (6.3% improvement).
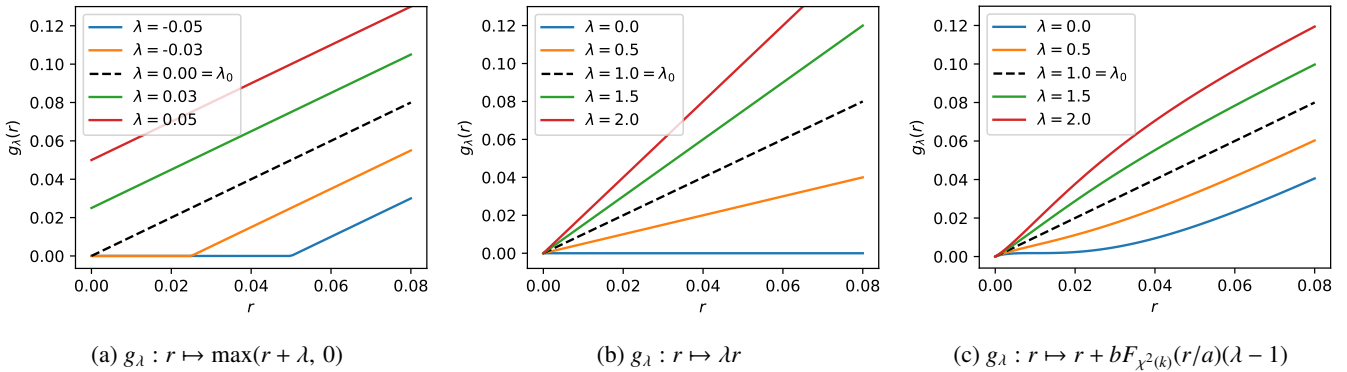
8

(a) $g_\lambda : r \mapsto \max(r + \lambda, 0)$ (b) $g_\lambda : r \mapsto \lambda r$ (c) $g_\lambda : r \mapsto r + b F_{\chi^2(k)}(r/a)(\lambda - 1)$

Fig. 4: Examples of families $(g_\lambda)_\lambda$ of calibration functions. In Fig. 4a (additive calibration), λ ranges from $-\infty$ to $+\infty$ with $\lambda_0 = 0$, whereas in Figs. 4b (multiplicative calibration) and 4c (chi-squared calibration), λ ranges from 0 to $+\infty$ with $\lambda_0 = 1$. In Fig. 4c, the number k of degrees of freedom is set to 3, the scaling factor a is set to 0.01, and the multiplicative factor b is set to the maximum value such that g_λ remains non-decreasing for any $\lambda \geq 0$, *i.e.*, $b \approx 0.041$.

Table 1: Reconstruction accuracy

| | RMSE ($\times 10^{-3}$) | | | |
Extent	KS1	KS2	Wiener	MCALens
All pixels	31.8 ± 0.5	21.1 ± 0.8	18.3 ± 1.2	$\mathbf{18.0 \pm 1.0}$
High-density	$\mathbf{60.1 \pm 2.6}$	66.8 ± 2.9	72.3 ± 3.0	67.7 ± 2.9

Note. The RMSE is computed between each ground truth convergence map κ_i and the corresponding reconstruction $\hat{\kappa}_i$. The empirical means and standard deviations are computed over the test set $(\gamma_i, \kappa_i)_{i=n+1}^m$ for each mass mapping method, and reported in this table. "KS1" and "KS2" correspond to the Kaiser-Squires estimators with weak and strong smoothing, respectively. The second row focuses on pixels k such that $|\kappa_i[k]| \geq 4.8 \times 10^{-2}$.

4.2.3. Miscoverage rate

For each $i \in \{n + 1, \ldots, m\}$ in the test set, we have measured the empirical miscoverage rate $L(\kappa_i, \hat{\kappa}_i^-, \hat{\kappa}_i^+)$ such as introduced in (35), where the (uncalibrated) lower- and upper-bounds $\hat{\kappa}_i^-$ and $\hat{\kappa}_i^+$ have been computed following (9), using each of the three reconstruction methods. Then, after calibrating the bounds using CQR and RCPS on the calibration set $(\gamma_i, \kappa_i)_{i=1}^n$, we have computed $L(\kappa_i, \hat{\kappa}_{\text{cqr}\, i}^-, \hat{\kappa}_{\text{cqr}\, i}^+)$ and $L(\kappa_i, \hat{\kappa}_{\text{rc}\, i}^-, \hat{\kappa}_{\text{rc}\, i}^+)$ on the test set $(\gamma_i, \kappa_i)_{i=n+1}^m$. Finally, we compared the results with the theoretical guarantees stated in (34) and (45), respectively. The results are reported in Table D.1 (Appendix D), and plotted in Fig. 5.

In the case of CQR, the empirical mean of the miscoverage rate falls between the theoretical bounds (yellow area). This is in line with the theoretical guarantee stated in (34). We also notice that the lower bound seems overly conservative. It should be noted that the expected value from (34) covers uncertainties over the convergence map **K**, the noise **N**, but also the calibration set $(\Gamma_i, \mathbf{K}_i)_{i=1}^n$. In our experiments, we only considered one realization $(\gamma_i, \kappa_i)_{i=1}^n$ of the calibration set, which, as explained in Sect. 3.4, may result in miscalibration and above-target error rates. This effect could be mitigated by increasing the size n of the calibration set, or with bootstrapping.

Unlike CQR, the RCPS approach controls the risk of picking a statistically-deviant calibration set, which could lead to undercoverage, by introducing an additional parameter δ. However, even with large values of δ, the calibrated bounds tend to be overly conservative. For example, with $\delta = 0.5$, one would ex-

pect the average miscoverage rate to fluctuate around the target α in approximately 50% of experiments when repeating the protocol with different calibration sets. In practice, the computed miscoverage rates consistently fall well below the target. Contrary to CQR, RCPS does not prevent situations of overcoverage.

4.2.4. Mean length of prediction intervals

The mean length of the prediction intervals has been computed over each image in the test set, before and after calibration, for each mass mapping method, each calibration method (CQR and RCPS), and each family of calibration functions. The results are reported in Table D.1 (Appendix D), and plotted in Fig. 6.

We observe that the choice of mass mapping method influences the size of the confidence intervals. Specifically, the KS solution yields larger calibrated error bars than either the Wiener or MCALens solutions, particularly in the weak-smoothing scenario ("KS1"). Additionally, the smallest confidence intervals are obtained with MCALens, using CQR with the additive family of calibration functions (in bold in Table D.1). This family consistently produces equal or smaller confidence intervals compared to the multiplicative or chi-squared families, for both CQR and RCPS calibration procedures. The modest improvement of the MCALens solution compared to Wiener (0.24% reduction in the mean length) can be attributed to the overall similarity between MCALens and Wiener outputs, with MCALens demonstrating higher reconstruction accuracy only in a few peak-like structures, which are essential for inferring cosmological parameters. Consequently, averaging the results over all pixels tends to hide essential properties of MCALens. A more detailed discussion on high-density regions is provided in Sect. 5.3.

As explained above, RCPS produces overly conservative bounds, which has a strong impact on the size of the confidence intervals. Smaller confidence intervals could be obtained by increasing the size n of the calibration set, since the Hoeffding's upper-confidence bound (38) decreases with increased values of n. Therefore, CQR seems more appropriate if the computational resources or available data are limited.

5. Discussion

5.1. Distribution-free vs Bayesian uncertainties

Extending distribution-free UQ to data-driven mass mapping methods (see Sect. 2.1) is straightforward, provided one has ac-
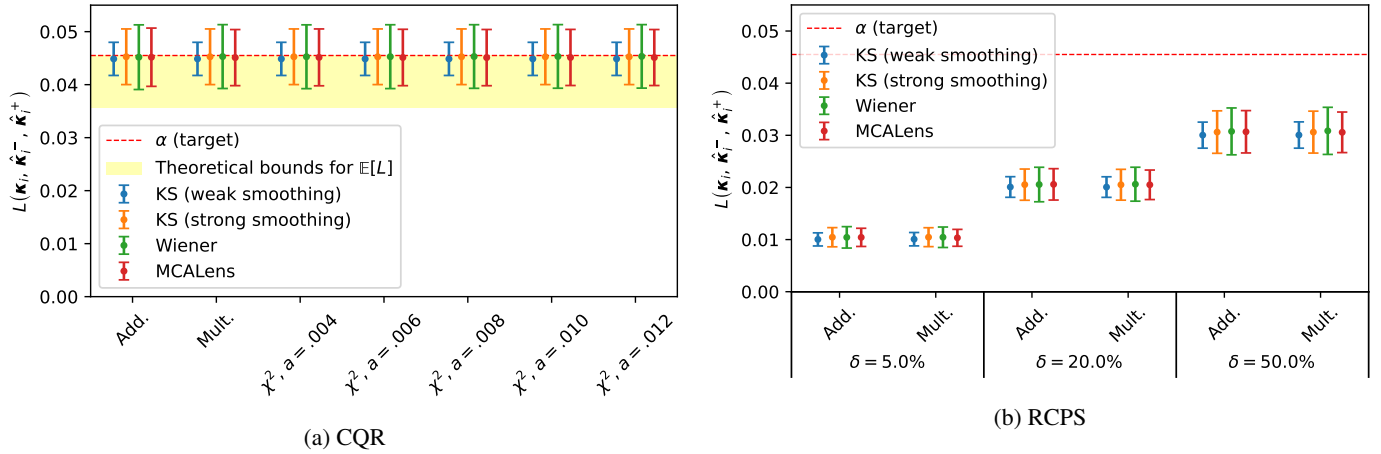
9

(a) CQR

(b) RCPS

Fig. 5: Empirical miscoverage rate after calibration with CQR (Fig. 5a) and RCPS (Fig. 5b). The means and standard deviations are computed over the test set $(\gamma_i, \kappa_i)_{i=n+1}^m$, for various error levels δ (RCPS only), and various families of calibration functions $(g_\lambda)_\lambda$. More details are provided in the note of Table D.1 (Appendix D). The theoretical bounds for CQR, introduced in (34), are represented by the yellow area in Fig. 5a. These plots indicate that CQR achieves miscoverage rates that are, on average, close to the target α, whereas RCPS tends to produce overconservative bounds, even for large values of δ.

(a) CQR

(b) RCPS
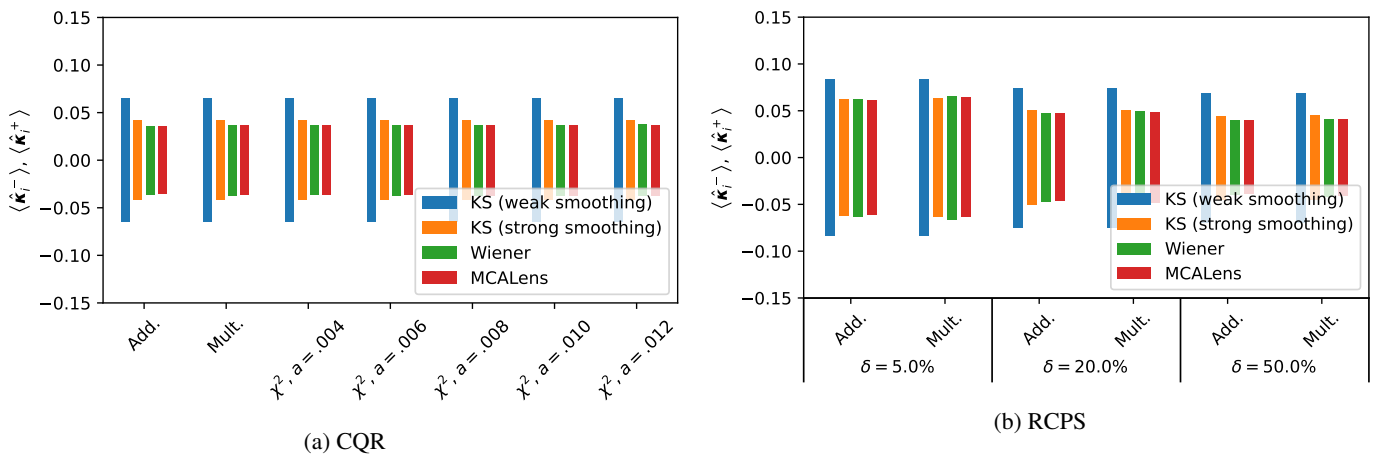
Fig. 6: Mean values for the lower and upper confidence bounds, computed over the span of each image. The bar sizes represent the mean length of the error bars. As in Fig. 5, the values are computed over the test set $(\gamma_i, \kappa_i)_{i=n+1}^m$, for various families of calibration functions $(g_\lambda)_\lambda$. After calibration with CQR, the choice of mass mapping method does not influence the average miscoverage rate (Fig. 5a), but significantly affects the size of the error bars (Fig. 6a). Additionally, RCPS, which tends to produce overconservative prediction bounds (Fig. 5b), yields larger error bars than CQR (Fig. 6b vs 6a).

cess to a "first guess", an initial quantification of uncertainty. Such initial uncertainty bounds can be obtained following the Bayesian framework presented in Sect. 2.2.2, which provides theoretical guarantees similar to (47).

However, as explained in Sect. 2.2.3, this heavily relies on a proper choice of the prior distribution μ, and if the latter differs from the unknown oracle distribution μ^*, the empirical mean of the miscoverage rate, computed over the test set, may significantly diverge from (47). Applying a post-processing calibration step, such as CQR or RCPS, allows for obtaining the guarantees stated in (34) or (45), respectively. In these expressions, the expected value assumes $\mathbf{K} \sim \mu^*$, even though the underlying distribution μ^* remains implicit—hence the term "distribution-free". The only sufficient conditions for these properties to be satisfied are that the data from the calibration and test sets $(\boldsymbol{\Gamma}_i, \mathbf{K}_i)_{i=1}^m$ are exchangeable, and that the conformity scores are almost surely distinct.

5.2. Beyond pixelwise uncertainty

This paper focuses on estimating error bars per pixel. While this approach is easily interpretable, it lacks some essential properties. First, it is important to consider the purpose of mass mapping and the type of information required. For instance, to infer cosmological parameters, several studies have proposed using peak counts (Kratochvil et al. 2010; Marian et al. 2013; Shan et al. 2014; Liu et al. 2015), convolutional neural networks that exploit information in the gradient around peaks (Ribli et al. 2019a,b), or neural compression of peak summary statistics (Jeffrey et al. 2021). Consequently, regions of higher density are of particular interest and should be treated with special attention.

In this context, alternative UQ schemes could be considered. Among these are the Bayesian hypothesis testing of structure (Cai et al. 2018a,b; Repetti et al. 2019; Price et al. 2021), which can assess whether a given structure in an image is a reconstruction artifact or has some physical significance. Additionally, a

computationally efficient approach, based on wavelet decomposition and thresholding of the MAP reconstruction, has been recently proposed by Liaudat et al. (2023). This method estimates errors at various scales, highlighting the different structures of the reconstructed image. In fact, these ideas can be traced back to earlier work in the stochastic geometry literature (see, *e.g.*, the book by Adler & Taylor 2007), with applications in neuroimaging (*e.g.*, Fadili & Bullmore 2004). Another potential direction is the application of calibration procedures to blob detection (Lindeberg 1993), for which UQ methods have been recently developed by Parzer et al. (2023, 2024).

Another drawback of pixelwise UQ is that it does not account for correlations between pixels. For instance, whether a given pixel has been accurately predicted can impact the uncertainty bounds of the neighboring pixels, a property not reflected in per-pixel error bars. This concern intersects with the previous one, as it raises the question of correctly identifying the structures of interest in the reconstructed image, which typically span several pixels. To address this issue, Belhasin et al. (2024) proposed applying UQ after decomposing the reconstructed images through principal component analysis. They also applied RCPS in this context, demonstrating that calibration methods can be used beyond the per-pixel framework.

5.3. Focus on higher-density regions

As confirmed by our experiments, CQR provides guarantees on the miscoverage rate $L(\kappa, \hat{\kappa}^-, \hat{\kappa}^+)$ defined in (35). However, this score is computed by considering all active pixels $k \in \mathcal{M}$. This may hide disparities caused by latent factors such as the local density of the convergence field. To support this claim, we computed the miscoverage rate filtered on regions of higher density. More precisely, for any $i \in \{n + 1, \ldots, m\}$, we only considered pixels k such that $|\kappa_i[k]| \geq t$ for a given threshold $t \in \mathbb{R}$, set to 4.8×10^{-2} in our experiments (S/N ratio above 0.25), similarly to Sect. 4.2.2. The corresponding metric is defined by

$$L_t(\kappa, \hat{\kappa}^-, \hat{\kappa}^+) := \frac{\operatorname{card}\left\{k \in \mathcal{M}_t(\kappa) : \kappa[k] \notin [\hat{\kappa}^-[k], \hat{\kappa}^+[k]]\right\}}{\operatorname{card} \mathcal{M}_t(\kappa)}, \quad (52)$$

with

$$\mathcal{M}_t(\kappa) := \left\{k \in \mathcal{M} : \kappa[k] \geq t\right\}. \quad (53)$$

The results, displayed in Fig. 7, indicate that the filtered miscoverage rate remains well above the target α, even after calibration with CQR. This phenomenon can also be visualized in Fig. 8 for the Wiener method: the bright area remains underestimated.

As highlighted in Sect. 5.2, these high-density regions are important for inferring cosmological parameters. Therefore, the method could be improved by incorporating this additional constraint. An interesting direction involves conformal prediction masks, as introduced by Kutiel et al. (2023). The method consists of masking the regions of low uncertainty in a given image, allowing a focus on the regions of higher uncertainty—corresponding to high-density regions in the case of weak lensing mass mapping.

Fig. 9 presents a scatter plot summarizing the various metrics discussed in this study. It indicates that Wiener and MCALens show similar error bar size and RMSE when computed across the entire set $\mathcal{M}$ of active pixels. From these metrics' perspective, both methods outperform the KS estimators. However, when focusing on high-density regions, the Wiener solution exhibits poorer performance in terms of reconstruction accuracy
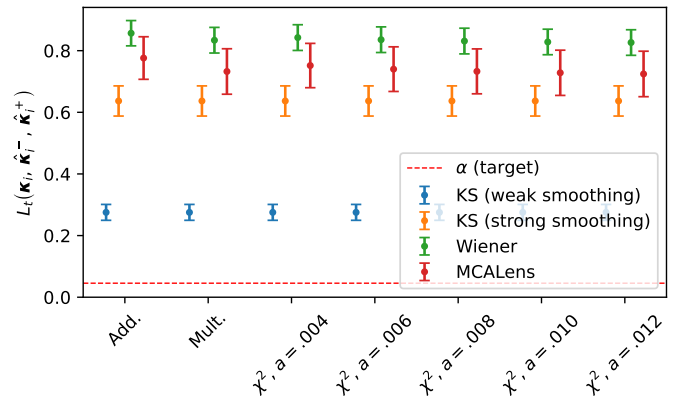

Fig. 7: Empirical miscoverage rate after calibration with CQR, computed on regions of higher density. Following Sect. 4.2.2, only pixels k with $|\kappa_i[k]| \geq t = 4.8 \times 10^{-2}$ were considered. Unlike in Fig. 5a, where the miscoverage rate is computed over all pixels, the measured values in this case deviate drastically from the target. Consequently, the size of the error bars is underestimated in these areas.
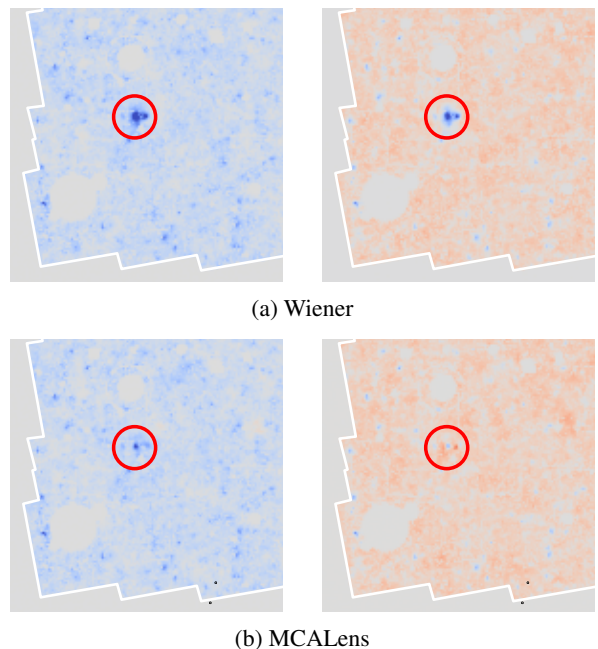

(a) Wiener

(b) MCALens

Fig. 8: Lower and upper bounds after calibration with CQR, with a focus on a peak-like structure. See Fig. D.1 for more details. In the Wiener solution, the ground truth $\kappa[k]$ remains above $\hat{\kappa}^+[k]$ (miscoverage), whereas MCALens provides better coverage. This observation aligns with Fig. 7.

(RMSE) and miscoverage rate, compared to MCALens and KS. As a result, MCALens achieves a balance between overall and high-density-specific performance. Additionally, the choice of smoothing filter significantly influences the reconstruction accuracy and error bar size of the KS estimator.

6. Conclusion

In this work, we emphasized the need for adjusting uncertainty estimates for reconstructed convergence maps, whether computed through frequentist or Bayesian frameworks. To ad-
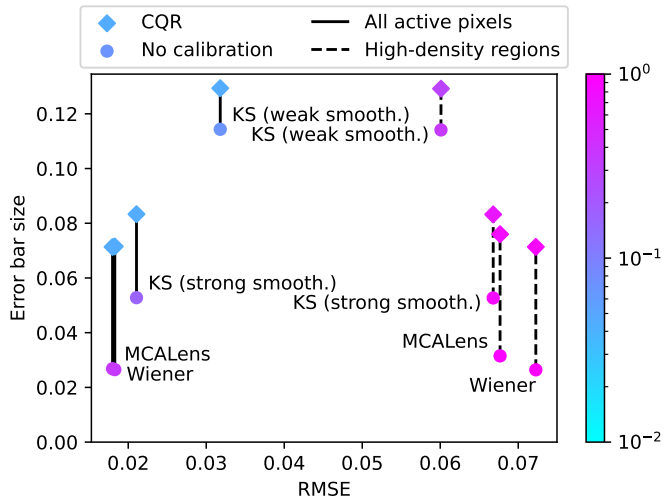
11

Fig. 9: Mean length of prediction intervals (Sect. 4.2.4) plotted against RMSE (Sect. 4.2.2), both before and after calibration using additive CQR, with and without filtering on high-density regions. Marker colors indicate empirical miscoverage rates (Sect. 4.2.3). The coverage guarantee (34) applies only after calibration (blue diamonds atop the solid lines). We notice that, when filtering on high-density regions, the coverage property no longer holds (pink diamonds atop the dashed lines).

dress this, we built on two recent distribution-free calibration methods—conformalized quantile regression (CQR) and risk-controlling prediction sets (RCPS)—to obtain error bars with valid coverage guarantees.

This paper presents several key innovations. First, while RCPS has been applied to inverse problems in other contexts, CQR was originally designed for scalar regression and required adaptation for this specific application. Second, no prior work has compared the two methods directly. Finally, we investigated various families of calibration functions for both approaches.

Our experiments led to three key findings. First, RCPS tends to produce overconservative confidence intervals, whereas CQR allows for more accurate—and smaller—error bars. Second, the choice of mass mapping method significantly influences the size of these error bars: a 45% reduction for MCALens compared to KS with weak smoothing, a 14% reduction compared to KS with strong smoothing, and a 0.24% reduction compared to Wiener. Finally, this choice also affects the reconstruction accuracy, especially around peak-like structures, where MCALens outperforms Wiener by 6.3%.

We emphasize that these calibration approaches are not limited to the model-driven mass mapping methods discussed in this paper; they are applicable to any mass mapping method, including those based on deep learning.

To conclude and guide future research, we acknowledge several limitations of this study. First, particular attention must be given to the intensity of peak-like structures, which are essential for inferring cosmological parameters. As it stands, neither of the two calibration methods effectively prevents miscoverage in these high-density regions. Additionally, this work could be expanded to include a broader range of uncertainty estimates beyond per-pixel error bars. Finally, the calibration and resulting uncertainty bounds were derived for a specific set of cosmological parameters. While we postulate that variations in cosmology have only a marginal impact on the results, this hypothesis needs further investigation.

Data availability

To ensure reproducibility, all software, scripts, and notebooks used in this study are available on GitHub.[4]

References

Adler, R. J. & Taylor, J. E. 2007, Random Fields and Geometry, Springer Monographs in Mathematics (Springer)

Angelopoulos, A. N., Bates, S., Malik, J., & Jordan, M. I. 2022a, in International Conference on Learning Representations

Angelopoulos, A. N., Kohli, A. P., Bates, S., et al. 2022b, in Proceedings of the 39th International Conference on Machine Learning (PMLR), 717–730

Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. 2021, J. ACM, 68, 43:1

Belhasin, O., Romano, Y., Freedman, D., Rivlin, E., & Elad, M. 2024, IEEE Trans. Pattern Anal. Mach. Intell., 46, 3321

Bobin, J., Starck, J.-L., Sureau, F., & Fadili, J. 2012, Adv. Astron., 2012, e703217

Boucheron, S., Lugosi, G., & Bousquet, O. 2004, in Advanced Lectures on Machine Learning (Springer), 208–240

Cai, X., Pereyra, M., & McEwen, J. D. 2018a, MNRAS, 480, 4154

Cai, X., Pereyra, M., & McEwen, J. D. 2018b, MNRAS, 480, 4170

Durmus, A., Moulines, É., & Pereyra, M. 2018, SIAM J. Imaging Sci., 11, 473

Fadili, M. J. & Bullmore, E. T. 2004, NeuroImage, 23, 1112

Horwitz, E. & Hoshen, Y. 2022, Conffusion: Confidence Intervals for Diffusion Models, arXiv:2211.09795

Jeffrey, N., Alsing, J., & Lanusse, F. 2021, MNRAS, 501, 954

Jeffrey, N., Lanusse, F., Lahav, O., & Starck, J.-L. 2020, MNRAS, 492, 5023

Jeffrey, N. & Wandelt, B. D. 2020, in Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020)

Kaiser, N. & Squires, G. 1993, ApJ, 404, 441

Kilbinger, M. 2015, Rep. Prog. Phys., 78, 086901

Klatzer, T., Dobson, P., Altmann, Y., et al. 2024, SIAM J. Imaging Sci., 1078

Kratochvil, J. M., Haiman, Z., & May, M. 2010, Phys. Rev. D, 81, 043519

Kutiel, G., Cohen, R., Elad, M., Freedman, D., & Rivlin, E. 2023, in ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare

Lanusse, F., Starck, J.-L., Leonard, A., & Pires, S. 2016, A&A, 591, A2

Laumont, R., Bortoli, V. D., Almansa, A., et al. 2022, SIAM J. Imaging Sci., 15, 701

Lei, J. & Wasserman, L. 2014, J. R. Stat. Soc., B: Stat. Methodol., 76, 71

Liaudat, T. I., Mars, M., Price, M. A., et al. 2023, Scalable Bayesian Uncertainty Quantification with Data-Driven Priors for Radio Interferometric Imaging, arXiv:2312.00125

Lindeberg, T. 1993, Int. J. Comput. Vis., 11, 283

Liu, J., Petri, A., Haiman, Z., et al. 2015, Phys. Rev. D, 91, 063507

Marian, L., Smith, R. E., Hilbert, S., & Schneider, P. 2013, MNRAS, 432, 1338

McEwen, J. D., Liaudat, T. I., Price, M. A., Cai, X., & Pereyra, M. 2023, Proximal Nested Sampling with Data-Driven Priors for Physical Scientists, arXiv:2307.00056

Mobasher, B., Capak, P., Scoville, N. Z., et al. 2007, The ApJ Supplement Series, 172, 117

Narnhofer, D., Habring, A., Holler, M., & Pock, T. 2024, SIAM J. Imaging Sci., 17, 301

Osato, K., Liu, J., & Haiman, Z. 2021, MNRAS, 502, 5593

Parzer, F., Jethwa, P., Boecker, A., et al. 2023, A&A, 674, A59

Parzer, F., Kirisits, C., & Scherzer, O. 2024, J. Math. Imaging Vis., 66, 697

Pereyra, M. 2016, Stat. Comput., 26, 745

Pereyra, M. 2017, SIAM J. Imaging Sci., 10, 285

Pereyra, M., Mieles, L. V., & Zygalakis, K. C. 2020, SIAM J. Imaging Sci., 13, 905

Price, M. A., Cai, X., McEwen, J. D., et al. 2020, MNRAS, 492, 394

Price, M. A., McEwen, J. D., Cai, X., et al. 2021, MNRAS, 506, 3678

Remy, B., Lanusse, F., Jeffrey, N., et al. 2023, A&A, 672, A51

Repetti, A., Pereyra, M., & Wiaux, Y. 2019, SIAM J. Imaging Sci., 12, 87

Ribli, D., Pataki, B. Á., & Csabai, I. 2019a, Nat. Astron., 3, 93

Ribli, D., Pataki, B. Á., Zorrilla Matilla, J. M., et al. 2019b, MNRAS, 490, 1843

Romano, Y., Patterson, E., & Candes, E. 2019, in Advances in Neural Information Processing Systems, Vol. 32 (Curran Associates, Inc.)

[4] https://github.com/hubert-leterme/weaklensing_uq.git

Romano, Y., Sesia, M., & Candes, E. 2020, in Advances in Neural Information Processing Systems, Vol. 33 (Curran Associates, Inc.)

Sadinle, M., Lei, J., & Wasserman, L. 2019, J. Am. Stat. Assoc., 114, 223

Schrabback, T., Hartlap, J., Joachimi, B., et al. 2010, A&A, 516, A63

Scoville, N., Abraham, R. G., Aussel, H., et al. 2007, The ApJ Supplement Series, 172, 38

Shan, H., Kneib, J.-P., Comparat, J., et al. 2014, MNRAS, 442, 2534

Shirasaki, M., Moriwaki, K., Oogi, T., et al. 2021, MNRAS, 504, 1825

Shirasaki, M., Yoshida, N., & Ikeda, S. 2019, Phys. Rev. D, 100, 043527

Starck, J.-L., Donoho, D. L., Fadili, M. J., & Rassat, A. 2013, A&A, 552, A133

Starck, J.-L., Elad, M., & Donoho, D. 2005, IEEE Trans. Image Process., 14, 1570

Starck, J.-L., Fadili, J., & Murtagh, F. 2007, IEEE Trans. Image Process., 16, 297

Starck, J.-L., Themelis, K. E., Jeffrey, N., Peel, A., & Lanusse, F. 2021, A&A, 649, A99

Steinwart, I. & Christmann, A. 2011, Bernoulli, 17, 211

Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. 2006, J. Mach. Learn. Res., 7, 1231

Teneggi, J., Tivnan, M., Stayman, W., & Sulam, J. 2023, in Proceedings of the 40th International Conference on Machine Learning (PMLR), 33940–33960

Vovk, V., Gammerman, A., & Shafer, G. 2005, Algorithmic Learning in a Random World (Springer-Verlag)

Appendix A: Mass mapping: problem formulation

The following formalism has been outlined by Kilbinger (2015).

A.1. Weak lensing and mass mapping

Consider a light ray emitted by an extended source (typically, a galaxy), observed at coordinates $\theta \in \mathbb{R}^2$ in the sky. In the presence of (inhomogeneous) matter density that we seek to estimate, the light ray undergoes continuous deflection as it travels through space, a phenomenon known as gravitational lensing. In the absence of lensing, it would be seen by the observer at coordinates $\beta(\theta)$. Assuming deflections are small enough (weak lensing regime), one can write

$$\beta(\theta) = \theta - \nabla\psi(\theta), \tag{A.1}$$

where $\psi : \mathbb{R}^2 \to \mathbb{R}$ denotes a lensing potential. Now, consider another light ray emitted by the same galaxy, observed at coordinates $\theta + \delta\theta$. A first-order approximation yields

$$\beta(\theta + \delta\theta) - \beta(\theta) = \mathbf{J}(\theta) \cdot \delta\theta, \tag{A.2}$$

where the Jacobian $\mathbf{J} : \mathbb{R}^2 \to \mathbb{R}^{2\times 2}$ (also referred to as amplification matrix) satisfies

$$\mathbf{J} := \begin{pmatrix} 1 - \kappa - \gamma_1 & -\gamma_2 \\ -\gamma_2 & 1 - \kappa + \gamma_1 \end{pmatrix} = (1-\kappa)\begin{pmatrix} 1 - g_1 & -g_2 \\ -g_2 & 1 - g_1 \end{pmatrix}, \tag{A.3}$$

where we have introduced the fields κ, γ_1, $\gamma_2 : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\kappa := \tfrac{1}{2}(\partial_1^2\psi + \partial_2^2\psi); \quad \gamma_1 := \tfrac{1}{2}(\partial_1^2\psi - \partial_2^2\psi); \quad \gamma_2 := \partial_1\partial_2\psi, \tag{A.4}$$

and $g_i := \gamma_i/(1+\kappa)$, $i = 1, 2$. For any $\theta \in \mathbb{R}^2$, $\kappa(\theta) \in \mathbb{R}$, referred to as the convergence, translates into an isotropic dilation of the source, and $\gamma(\theta) := \gamma_1(\theta) + i\gamma_2(\theta) \in \mathbb{C}$, referred to as the shear, causes anisotropic stretching of the image, typically transforming a circle into an ellipse. Finally, $g(\theta) := g_1(\theta) + ig_2(\theta)$ is referred to as the reduced shear.

The goal is now to estimate the projected mass between the observer and the observed galaxies in any direction. In the weak lensing regime, it can be shown that the convergence $\kappa(\theta)$ is roughly proportional to the (weighted) matter density projected along the line of sight between the observer and the source. Therefore, mass mapping can be performed by estimating the convergence map $\kappa : \mathbb{R}^2 \to \mathbb{R}$ from the observation of galaxies at a given redshift.

A.2. Relationship between shear and convergence

Although not directly measurable, κ can be retrieved from the shear γ, up to an additive constant. In the flat sky limit, solving the PDEs (A.4) yields the following relationship in the Fourier space:

$$\forall \nu \neq \mathbf{0}, \ \tilde{\gamma}(\nu) = \frac{1}{\pi}\tilde{\mathcal{D}}(\nu)\tilde{\kappa}(\nu), \ \text{with } \tilde{\mathcal{D}} : \nu \mapsto \pi\frac{(\nu_1 + i\nu_2)^2}{\nu_1^2 + \nu_2^2}. \tag{A.5}$$

Denoting by $\mathcal{F}$ the Fourier transform operator, this yields

$$\gamma(\theta) = \frac{1}{\pi}\mathcal{F}^{-1}\big(\tilde{\mathcal{D}}\tilde{\kappa}\big)(\theta). \tag{A.6}$$

14

Now, we discretize (A.6) on a square grid of size $K^2 \in \mathbb{N}$, and adopt a vector-matrix notation in which the shear and convergence maps $\gamma \in \mathbb{C}^{K^2}$ and $\kappa \in \mathbb{R}^{K^2}$ are flattened and thus represented as one-dimensional vectors. By neglecting discretization and boundary effects, we get the following relation:

$$\gamma = \mathbf{A}\kappa + \gamma_0, \qquad \text{with} \qquad \mathbf{A} := \mathbf{F}^*\mathbf{P}\mathbf{F}, \tag{A.7}$$

where $\mathbf{F}$ and its Hermitian transpose $\mathbf{F}^* \in \mathbb{C}^{K^2\times K^2}$ respectively encode the discrete Fourier and inverse Fourier transforms, and $\mathbf{P} \in \mathbb{C}^{K^2\times K^2}$ is a diagonal matrix obtained by discretizing $\tilde{\mathcal{D}}/\pi$ on the same grid: for any $(k_1, k_2) \neq (0, 0)$,

$$\mathbf{P}\big[K^2 k_1 + k_2, \ K^2 k_1 + k_2\big] := \frac{(k_1 + ik_2)^2}{k_1^2 + k_2^2}. \tag{A.8}$$

The constant $\gamma_0 := \langle\gamma\rangle \in \mathbb{C}$ (mean value of the shear map) accounts for the fact that $\tilde{\mathcal{D}}$ is discontinuous in $\mathbf{0}$, thereby leaving $\mathbf{P}[0, 0]$ undefined when discretizing (mass-sheet degeneracy). In practice, we therefore set $\mathbf{P}[0, 0] = 0$ and add this constant to the equation.

Due to mass-sheet degeneracy, the operator $\mathbf{A}$ is not invertible. Therefore, assuming the shear map γ is known, the convergence map κ can only be retrieved up to an additive constant:

$$\kappa = \mathbf{A}^\dagger\gamma + \kappa_0, \qquad \text{with} \qquad \mathbf{A}^\dagger := \mathbf{F}^*\mathbf{P}^\dagger\mathbf{F}, \tag{A.9}$$

where $\kappa_0 := \langle\kappa\rangle \in \mathbb{R}$ denotes the mean value of the convergence map, and $\mathbf{P}^\dagger \in \mathbb{C}^{K^2\times K^2}$ denotes a diagonal matrix defined by $\mathbf{P}^\dagger[0, 0] := 0$ and, for all $k \neq 0$, by $\mathbf{P}^\dagger[k, k] := \mathbf{P}[k, k]^{-1}$. In practice, we will therefore estimate the variations of κ around its mean value.

A.3. Shear measurement

The intrinsic ellipticity of a galaxy is characterized by a complex number $\epsilon^s \in \mathbb{C}$. In the weak lensing regime, the observed ellipticity ϵ approximately satisfies

$$\epsilon = g + \epsilon^s, \tag{A.10}$$

where, as a reminder, $g \in \mathbb{C}$ denotes the reduced shear, which varies with the galaxy coordinates θ. A common approximation in weak lensing is to consider $\kappa \ll 1$, and therefore $g \approx \gamma$. In the remaining of the section, we will replace g by the unreduced shear γ. Assuming the intrinsic galaxy ellipticity has no preferred orientation, the expected value of ϵ^s vanishes, which yields $\mathbb{E}[\epsilon] = \mathbb{E}[\gamma]$. Therefore, the observed ellipticity ϵ is an unbiased estimator of γ.

In practice, we consider a portion of the sky on which one seek to perform mass mapping, which we subdivide into K^2 small regions (or bins), each of which associated to a pixel in the discretized shear map γ and convergence map κ. For each bin k, we measure the ellipticity of N_k galaxies at a given redshift. Then, (A.10) yields

$$\frac{1}{N_k}\sum_{i=1}^{N_k}\epsilon_{k,i} = \frac{1}{N_k}\sum_{i=1}^{N_k}\gamma_{k,i} + \frac{1}{N_k}\sum_{i=1}^{N_k}\epsilon_{k,i}^s, \tag{A.11}$$

where $\gamma_{k,i}$ denotes the shear, and $\epsilon_{k,i}$ and $\epsilon_{k,i}^s$ respectively denote the intrinsic and measured ellipticity of the i-th galaxy within the k-th bin. Assuming that the bins are small enough to maintain an approximate consistancy of the shear within each, we get

$$\frac{1}{N_k}\sum_{i=1}^{N_k}\gamma_{k,i} = \gamma[k]. \tag{A.12}$$

Moreover, we introduce

$$\epsilon[k] := \frac{1}{N_k} \sum_{i=1}^{N_k} \epsilon_{k,i} \quad \text{and} \quad n[k] := \frac{1}{N_k} \sum_{i=1}^{N_k} \epsilon_{k,i}^s, \tag{A.13}$$

in which ϵ is an unbiased estimator of γ, and n is the realization of a Gaussian noise with zero mean and diagonal covariance matrix $\Sigma_n \in \mathbb{C}^{K^2 \times K^2}$, satisfying

$$\Sigma_n[k, k] = \sigma_{\text{ell}}/\sqrt{N_k}, \tag{A.14}$$

where $\sigma_{\text{ell}} \in \mathbb{C}$ denotes the standard deviation of the intrinsic ellipticities. In practice, this quantity is estimated by measuring the standard deviation of the observed ellipticities across the sky. In practice, the number of galaxies per bin may be quite small ($N_k \sim 10$), which yields a very low S/N ratio.

Then, combining (A.7) and (A.11)-(A.13) yields

$$\epsilon = \mathbf{A}\kappa + n + \gamma_0. \tag{A.15}$$

Since the noise n has zero-mean, we have $\gamma_0 \approx \langle \epsilon \rangle$. Therefore, by replacing the true unknown shear map by its unbiased estimator, and after applying mean-centering: $\gamma := \epsilon - \langle \epsilon \rangle$, we get the inverse problem stated in (1), for which we seek an estimation of the convergence map κ.

A.4. Masked data

In practice, some data may be missing due to survey measurement masks (for instance, limits of the galaxy survey or bright stars in the foreground). We consider the set of "active" pixels, denoted by

$$\mathcal{M} := \left\{ k \in \{1, \dots, K^2\} : N_k > 0 \right\}. \tag{A.16}$$

A naive approach consists in setting to $\gamma[k] = 0$ for any $k \notin \mathcal{M}$. Then, if we assign $n[k]$ to $-\mathbf{A}\kappa[k]$ outside $\mathcal{M}$, (1) still holds, but n can no longer be considered as a Gaussian noise. For the sake of simplicity, in this paper, we nevertheless relied on this solution, and assigned a very large noise variance $\Sigma_n[k, k]$ for any $k \notin \mathcal{M}$, which is required for the Wiener and MCALens mass mapping methods (see Sects. B.2 and B.3, respectively).

Another approach involves adding a Gaussian white noise n' with large variance σ'^2 to the masked pixels: $\gamma[k] := n'[k]$ for any $k \notin \mathcal{M}$. Assuming the S/N ratio between $\mathbf{A}\kappa$ and n' is small enough, we therefore approximately get (1), where the noise variance is left unchanged for unmasked pixels, and is equal to σ'^2 in masked regions.

Because of noise and masked data, this inverse problem is ill-posed. Therefore, mass-mapping methods generally rely on prior assumptions to regularize the problem.

Appendix B: Mass mapping methods

In this section, we present the three mass mapping methods used in our experiments.

B.1. Kaiser-Squires inversion

The Kaiser-Squires (KS) inversion (Kaiser & Squires 1993) consists in a simple pseudo-inversion of the linear operator $\mathbf{A}$. According to (A.9), in absence of noise and mask, it provides an exact reconstruction of the true convergence map κ (up to an additive constant). In practice, it is generally followed by a certain amount of Gaussian smoothing:

$$\hat{\kappa}_{\text{ks}} := \mathbf{S}\mathbf{A}^\dagger \gamma, \tag{B.1}$$

where $\mathbf{S} \in \mathbb{R}^{K^2 \times K^2}$ encodes a Gaussian smoothing operator with manually-tuned variance.

Note that, in absence of smoothing ($\mathbf{S} = \mathbf{I}$), the KS solution corresponds to the maximum likelihood estimate of the convergence map:

$$\hat{\kappa}_{\text{ks}} \in \text{argmin}_{\kappa'} \left\{ -\log f_{\Gamma|\kappa}(\gamma|\kappa') \right\} \tag{B.2}$$

$$= \text{argmin}_{\kappa'} \frac{1}{2} \left\| \gamma - \mathbf{A}\kappa' \right\|_{\Sigma_n^{-1}}^2, \tag{B.3}$$

where $f_{\Gamma|\kappa}(\gamma|\kappa')$ denotes the likelihood density.

B.2. Iterative Wiener filtering

The KS inversion generally produces poor results because it is very sensitive to noise and masked data. To overcome this limitation, one introduce prior assumptions in the form of a regularization term in the objective function. Considering κ as drawn from a random variable with probability density function p, the MAP estimate satisfies (19).

In the Wiener approach, we consider κ as a Gaussian field. That is, it is drawn from a multivariate Gaussian distribution with (non-diagonal) covariance matrix Σ_κ, and the covariance matrix of its Fourier transform $\tilde{\kappa}$, denoted by $\mathbf{P}_\kappa := \mathbf{F}\Sigma_\kappa \mathbf{F}^*$, is diagonal. Then, the distribution from which κ is drawn is entirely characterized by its power spectrum. In this context, (19) becomes

$$\hat{\kappa}_{\text{wien}} \in \text{argmin}_{\kappa'} \left\{ \frac{1}{2} \left\| \gamma - \mathbf{A}\kappa' \right\|_{\Sigma_n^{-1}}^2 + \frac{1}{2} \left\| \kappa' \right\|_{\Sigma_\kappa^{-1}}^2 \right\}. \tag{B.4}$$

By setting the gradient of the above expression to 0, we get

$$\hat{\kappa}_{\text{wien}} = \left(\mathbf{A}^* \Sigma_n^{-1} \mathbf{A} + \Sigma_\kappa^{-1} \right)^{-1} \mathbf{A}^* \Sigma_n^{-1} \gamma. \tag{B.5}$$

If the noise is not stationary, then computing (B.5) requires two $K^2 \times K^2$-matrix inversions, which is not feasible in practice. To overcome this, Bobin et al. (2012) proposed an iterative algorithm converging toward (B.5), in the context of cosmic microwave background (CMB) map restoration. Each iteration consists in a two-step process: one step in the spatial domain and one step in the Fourier domain. The algorithm takes advantage of the fact that the covariance matrices Σ_n and $\mathbf{P}_\kappa$ are diagonal, thereby avoiding large matrix inversions.

B.3. MCALens algorithm

In contrast to the CMB, matter distribution is actually poorly approximated by Gaussian fields, due to the gravitational interactions responsible for the formation of galaxies and large-scale structures. In a more recent paper, Starck et al. (2021) suggested that convergence maps can be represented as the superposition of a Gaussian component, as in the Wiener solution, and a sparse component in a wavelet dictionary, as in GLIMPSE2D:

$$\hat{\kappa}_{\text{mcal}} = \hat{\kappa}_{\text{G}} + \hat{\kappa}_{\text{S}}. \tag{B.6}$$

The authors then proposed to use the morphological component analysis (MCA) algorithm (Starck et al. 2005), considering that the Gaussian and sparse components are morphologically distinct (except for lower frequencies). They came along

with an iterative algorithm called MCALens, where each iteration is a two-step process. First, perform one iteration of the GLIMPSE2D algorithm on the residual $\gamma - \mathbf{A}\hat{\kappa}_G$, and update the sparse component $\hat{\kappa}_S$. Then, perform one iteration of Wiener filtering, described in Sect. B.2, on the residual $\gamma - \mathbf{A}\hat{\kappa}_S$, and update the Gaussian component $\hat{\kappa}_G$. Finally, since we consider the variations of the convergence around its mean value, the output of the MCALens algorithm undergoes mean-centering.

Appendix C: Proof of Proposition 1

Proof. The exchangeability of $(\mathsf{X}_i, \mathsf{Y}_i)_{i=1}^{n+1}$ implies the exchangeability of the conformity scores $(\Lambda_i)_{i=1}^{n+1}$. Moreover, by hypothesis, the latter are almost surely distinct. Therefore, we can apply Lemma 2 by Romano et al. (2019, supplementary material):

$$1 - \alpha \le \mathbb{P}\left\{\Lambda \le \Lambda^{(\alpha)}\right\} \le 1 - \alpha + \frac{1}{n+1}, \tag{C.1}$$

where we have denoted $\Lambda := \Lambda_{n+1}$. Let λ and $\lambda^{(\alpha)}$ denote realizations of Λ and $\Lambda^{(\alpha)}$, respectively. Then, λ corresponds to the (unknown) conformity score associated with the outcome (x, y) of (X, Y). It satisfies, similarly to (26)-(27),

$$\lambda = \min\left\{\lambda' \in [a, b] : y \in \hat{C}_{\lambda'}(x)\right\} \tag{C.2}$$

$$= \min\left\{\lambda' \in [a, b] : g_{\lambda'}(\hat{r}(x)) \ge \left|\hat{f}(x) - y\right|\right\}. \tag{C.3}$$

Then, combining (C.2), (C.3) and (24) yields, for any $\lambda' \in \mathbb{R}$,

$$\lambda \le \lambda' \iff g_{\lambda'}(\hat{r}(x)) \ge \left|\hat{f}(x) - y\right| \tag{C.4}$$

$$\iff y \in \hat{C}_{\lambda'}(x). \tag{C.5}$$

Therefore, using $\lambda' := \lambda^{(\alpha)}$, we get

$$\mathbb{P}\left\{\Lambda \le \Lambda^{(\alpha)}\right\} = \mathbb{P}\left\{\mathsf{Y} \in \hat{C}_{\Lambda^{(\alpha)}}(\mathsf{X})\right\}. \tag{C.6}$$

Finally, plugging (C.6) into (C.1) yields (29), which concludes the proof.

Appendix D: Detailed visualizations and results

Comprehensive visualizations of mass estimates and their uncertainties are shown in Fig. D.1, while detailed numerical results are provided in Table D.1.

Table D.1: UQ metrics: miscoverage rate and mean length of prediction intervals

Type of calibration		Miscoverage rate (%)				Mean length of pred. intervals ($\times 10^{-2}$)			
		KS1	KS2	Wiener	MCALens	KS1	KS2	Wiener	MCALens
Uncalibrated		6.94 ± 0.39	16.36 ± 1.09	36.04 ± 1.93	35.53 ± 1.89	11.43 ± 0.00	5.28 ± 0.00	2.65 ± 0.02	2.68 ± 0.02
CQR	Add.	4.49 ± 0.31	4.53 ± 0.53	4.52 ± 0.61	4.52 ± 0.55	12.94 ± 0.00	8.33 ± 0.00	7.15 ± 0.02	$\mathbf{7.13 \pm 0.02}$
	Mult.	4.49 ± 0.31	4.53 ± 0.53	4.53 ± 0.60	4.51 ± 0.53	12.94 ± 0.00	8.33 ± 0.00	7.39 ± 0.04	7.35 ± 0.06
	χ^2, $a = 0.004$	4.49 ± 0.31	4.53 ± 0.53	4.53 ± 0.60	4.51 ± 0.54	12.94 ± 0.00	8.33 ± 0.00	7.31 ± 0.04	7.26 ± 0.05
	χ^2, $a = 0.006$	4.49 ± 0.31	4.53 ± 0.53	4.53 ± 0.60	4.51 ± 0.53	12.94 ± 0.00	8.33 ± 0.00	7.38 ± 0.04	7.33 ± 0.06
	χ^2, $a = 0.008$	4.49 ± 0.31	4.53 ± 0.53	4.53 ± 0.60	4.51 ± 0.53	12.94 ± 0.00	8.33 ± 0.00	7.42 ± 0.05	7.37 ± 0.06
	χ^2, $a = 0.010$	4.49 ± 0.31	4.53 ± 0.53	4.53 ± 0.60	4.51 ± 0.53	12.94 ± 0.00	8.33 ± 0.00	7.45 ± 0.05	7.40 ± 0.07
	χ^2, $a = 0.012$	4.49 ± 0.31	4.53 ± 0.53	4.53 ± 0.60	4.51 ± 0.53	12.94 ± 0.00	8.33 ± 0.00	7.48 ± 0.05	7.42 ± 0.07
RCPS	$\delta = 0.05$ Add.	1.00 ± 0.13	1.05 ± 0.18	1.04 ± 0.21	1.04 ± 0.17	16.72 ± 0.00	12.45 ± 0.00	12.56 ± 0.02	12.26 ± 0.02
	$\delta = 0.05$ Mult.	1.01 ± 0.13	1.05 ± 0.18	1.04 ± 0.20	1.03 ± 0.16	16.76 ± 0.00	12.61 ± 0.00	13.21 ± 0.08	12.75 ± 0.12
	$\delta = 0.20$ Add.	2.01 ± 0.20	2.05 ± 0.30	2.06 ± 0.33	2.06 ± 0.30	14.87 ± 0.00	10.05 ± 0.00	9.47 ± 0.02	9.35 ± 0.02
	$\delta = 0.20$ Mult.	2.01 ± 0.20	2.05 ± 0.30	2.06 ± 0.33	2.05 ± 0.28	14.90 ± 0.00	10.15 ± 0.00	9.90 ± 0.06	9.71 ± 0.09
	$\delta = 0.50$ Add.	3.00 ± 0.25	3.06 ± 0.41	3.08 ± 0.45	3.07 ± 0.41	13.79 ± 0.00	8.96 ± 0.00	7.91 ± 0.02	7.88 ± 0.02
	$\delta = 0.50$ Mult.	3.01 ± 0.25	3.06 ± 0.40	3.09 ± 0.45	3.06 ± 0.39	13.80 ± 0.00	9.03 ± 0.00	8.30 ± 0.05	8.23 ± 0.07

Note. The metrics are computed for each mass mapping method, before and after calibration, for various error levels δ (RCPS only), and various families of calibration functions $(g_\lambda)_\lambda$. "Add." and "Mult." refer to the families defined in (49) and (50), respectively, whereas "χ^2" refers to the chi-squared family defined in (51), with $k = 3$ degrees of freedom. The multiplicative factor b, which depends on the scaling factor a, is set to the maximum value such that g_λ remains non-decreasing for any $\lambda \geq 0$. "KS1" and "KS2" correspond to the Kaiser-Squires estimators with weak and strong smoothing, respectively. The miscoverage rate (35) as well as the length of the prediction intervals are averaged over the set $\mathcal{M}$ of active pixels, for each pair $(\boldsymbol{\gamma}_i, \boldsymbol{\kappa}_i)$. The empirical mean and standard deviation of each metric is then computed over the test set $(\boldsymbol{\gamma}_i, \boldsymbol{\kappa}_i)_{i=n+1}^m$, and displayed in this table.
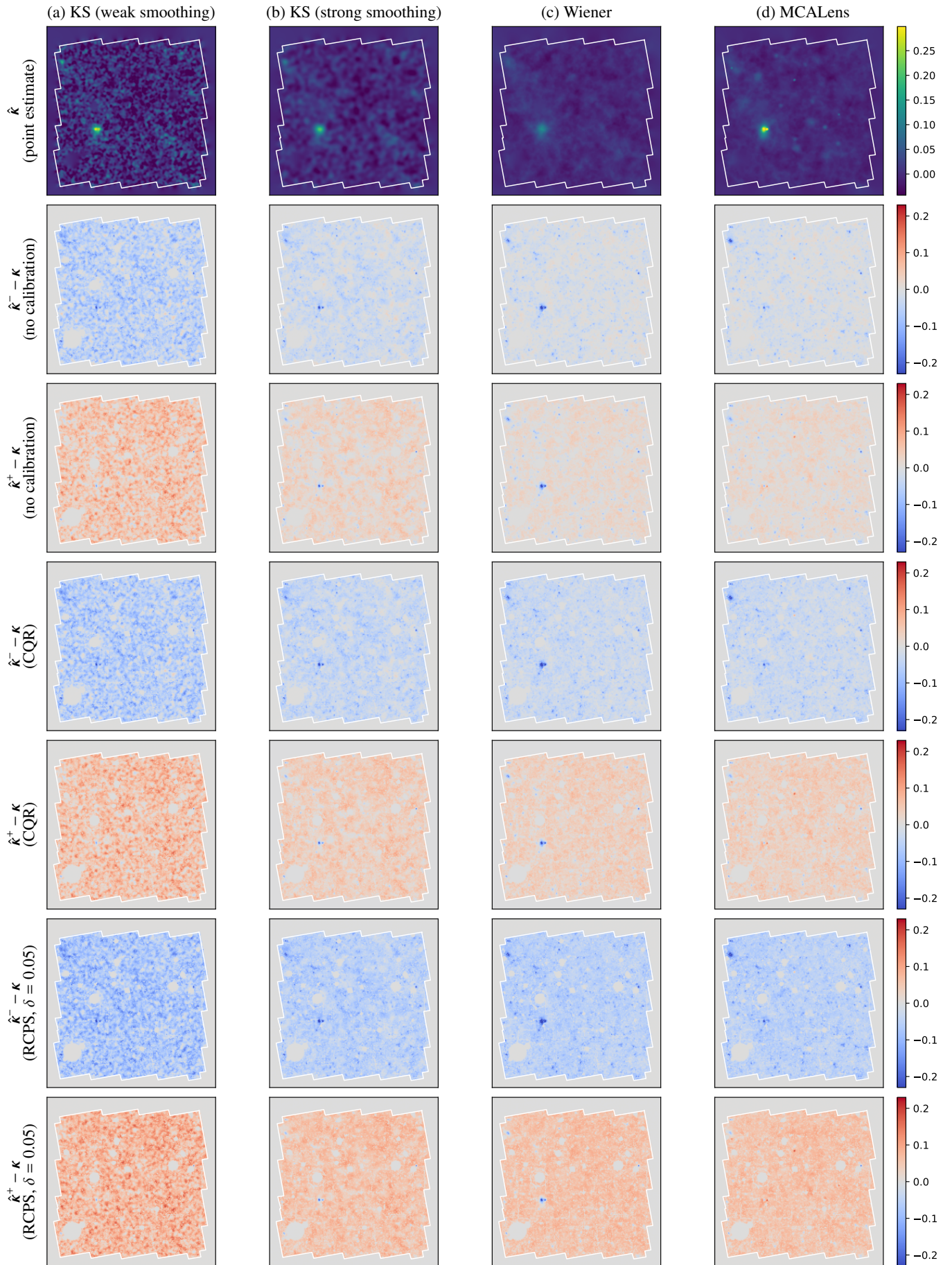
Fig. D.1: Top row: point estimate $\hat{\boldsymbol{\kappa}}$. Remaining rows: lower and upper bounds $\hat{\boldsymbol{\kappa}}^-$ and $\hat{\boldsymbol{\kappa}}^+$, before and after calibration (additive calibration), for each of the three mass mapping methods. The bounds have been centered with respect to the ground truth $\boldsymbol{\kappa}$.