Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **417,** 1938–1951 (2011) doi:10.1111/j.1365-2966.2011.19382.x

Probe combination in large galaxy surveys: application of Fisher information and Shannon entropy to weak lensing

J. Carron,* A. Amara and S. J. Lilly
Insitute of Astronomy, ETH Zuerich, Zurich CH-8093, Switzerland

Accepted 2011 July 4. Received 2011 June 21; in original form 2010 December 6

ABSTRACT

This paper aims at developing a better understanding of the structure of the information that is contained in galaxy surveys, so as to find optimal ways to combine observables from such surveys. We first show how Jaynes' Maximum Entropy Principle allows us, in the general case, to express the Fisher information content of data sets in terms of the curvature of the Shannon entropy surface with respect to the relevant observables. This allows us to understand the Fisher information content of a data set, once a physical model is specified, independently of the specific way that the data will be processed, and without any assumptions of Gaussianity. This includes as a special case the standard Fisher matrix prescriptions for Gaussian variables widely used in the cosmological community, for instance for power spectra extraction. As an application of this approach, we evaluate the prospects of a joint analysis of weak lensing tracers up to the second order in the shapes distortions, in the case that the noise in each probe can be effectively treated as model-independent. These include the magnification, and the two ellipticity and four flexion fields. At the two-point level, we show that the only effect of treating these observables in combination is a simple scale-dependent decrease in the noise contaminating the accessible spectrum of the lensing E-mode. We provide simple bounds to its extraction by a combination of such probes as well as its quantitative evaluation when the correlations between the noise variables for any two such probes can be ignored.

Key words: methods: data analysis – methods: statistical – cosmological parameters – large-scale structure of Universe.

1 INTRODUCTION

With cosmological data sets currently going through a rapid period of growth, it is increasingly important to quantitatively understand the potential and limits of particular data sets to test a physical model or hypothesis. For this, the Fisher information matrix has become a widely used tool in cosmology.

The concept of Fisher information has a long history. It was first coined by the statistician and geneticist R. A. Fisher (Fisher 1925) under the name of intrinsic accuracy of frequency curves. It has found its way into the cosmological community over the last decade, where it is often used to optimize survey configurations (Albrecht et al. 2006; Amara & Réfrégier 2007; Parkinson et al. 2007; Bernstein 2009) of planned cosmology experiments or to evaluate the expected errors on certain cosmological parameters with some observables (Tegmark 1997; Tegmark, Taylor & Heavens 1997; Hu & Tegmark 1999; Hu & Jain 2004).

Much of the work to date has been limited to particular sets of observables and estimators. Usually, it is assumed that observational errors as well as the parameters' probability distribution have Gaussian shape. The first aim of this work is to propose a framework to express the global Fisher information content of large data sets in a way that is independent of the specific ways that the data will be processed, and in realistic situations, where the exact statistical properties of the data are not known precisely. This should then provide a well-motivated basis point in order to perform systematic and robust trade-off studies. For this purpose a number of useful concepts already exist in the fields of information theory and probability theory, such as Shannon entropy or relative entropy (Kullback 1959), which we can use to gain a better understanding of what we can achieve with planned experiments. Specifically, we will show that we can achieve our aim by combining Fisher's information measure with Jaynes' Principle of Maximum Entropy (Jaynes 1983; Jaynes & Bretthorst 2003).

In a second step, as a concrete application of this approach, we investigate the joint entropy and information content of multiple observables of the same underlying, cosmologically interesting field. This is a very relevant situation in weak lensing (Schneider, Ehlers

*E-mail: jcarron@phys.ethz.ch

& Falco 1992; Bartelmann & Schneider 2001; Refregier 2003; Munshi et al. 2006; Schneider, Kochanek & Wambsganss 2006), where the distortions of galaxy images to any order are sourced by the lensing potential field.

This paper is divided into the following sections. In Section 2, we present our approach in details. We first review and develop some key properties of Fisher information, and its link to the Cramer–Rao inequality. We put a strong emphasis on its interpretation as a measure of the information on the model parameters in a data set, which is obtained from the probability distribution of different observational outcomes as a function of these same model parameters. Readers familiar with these aspects may jump to Section 2.3, where we introduce Jaynes' Maximum Entropy Principle, and show how it ideally completes Fisher's information measure, allowing us to understand the information content of a data set on a physical model in the case of incomplete knowledge. In Section 3, we show how the study of the Shannon entropy of a set of homogeneous fields provides a simple and model parameter independent answer to the question of the combination of the weak lensing observables' shear, magnification and flexion. We provide in Section 4 a quantitative evaluation of the prospects of such a combination at the two-point level for typical dark energy survey parameters, and conclude in Section 5 with a summary of the results and a discussion. A set of Appendices presents some technical details.

2 FISHER INFORMATION AND JAYNES' MAXENT PRINCIPLE

The concept of Fisher information is rich and not limited to parameter error estimation. We review here a few simple points of interest that justify the interpretation of the Fisher matrix as a measure of the information content of an experiment. Let us begin by considering the case of a single measurement X, with different possible outcomes, or realizations, x, and our model has a single parameter α. We also assume that we have knowledge, prior to the given experiment, of the probability density function $p_X(x, \alpha)$, which depends on our parameter α, which gives the probability of observing particular realizations for each value of the model parameter. The Fisher information, F, in X on α, is a non-negative scalar in this one-parameter case. It is defined in a fully general way as a sum over all realizations of the data (Fisher 1925):

$$F^X(\alpha) = \left\langle \left(\frac{\partial \ln p_X(x,\alpha)}{\partial \alpha} \right)^2 \right\rangle . \tag{1}$$

Angle brackets will always stand for mean value with respect to the probability density function, i.e. for any function f,

$$\langle f \rangle \equiv \int \mathrm{d}x \; p_X(x,\alpha) f(x). \tag{2}$$

Three simple but important properties of Fisher information are worth highlighting at this point.

The first is that $F^X(\alpha)$ is positive definite, and it vanishes if and only if the parameter α does not impact the data, i.e. if the derivative of $p_X(x, \alpha)$ with respect to α is zero for every realization x.

The second point is that it is invariant to invertible manipulations of the observed data. This can be seen by considering an invertible change of variable $y = f(x)$, which, due to the rules of probability theory, can be expressed as

$$p_Y(y,\alpha) = p_X(x,\alpha) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| . \tag{3}$$

Thus

$$\frac{\partial \ln p_Y(y,\alpha)}{\partial \alpha} = \frac{\partial \ln p_X(x,\alpha)}{\partial \alpha}, \tag{4}$$

leading to the simple equivalence that $F^X(\alpha) = F^Y(\alpha)$. On the other hand, information may be lost when the transformation is not unique in both the directions (e.g., see Rao 1973, for a proof). For instance, if the data are combined to produce a new variable that could arise from different sets of data points. This means manipulations of the data lead, at best, only to conservation of the information.

The third point is that information from independent experiments add together. Indeed, if two experiments with data X and Y are independent, then the joint probability density factorizes:

$$p_{XY}(x, y) = p_X(x) p_Y(y), \tag{5}$$

and it is easy to show that the joint information in the observations decouples:

$$F^{XY}(\alpha) = F^X(\alpha) + F^Y(\alpha). \tag{6}$$

These properties are making the Fisher information a meaningful measure of information. This is independent of its interpretation as providing error bars on parameters. It further implies that once a physical model is specified with a given set of parameters, a given experiment has a definite information content that can only decrease with data processing.

2.1 The case of a single observable

To quantify the last point above, and in order to understand the structure of the information in a data set, we first review a simple situation, common in cosmology, where the extraction of the model parameter α from the data goes through the intermediate step of estimating a particular observable, D, from the data, x, with the help of which α will be inferred. A typical example could be, from the temperature map of the CMB (x), the measurement of the power spectra of the fluctuations (D), from which a cosmological parameter (α) is extracted. The observable D is measured from x with the help of an estimator, which we call $\hat{D}$ and which we will take as unbiased. This means that its mean value, as would be obtained for instance if many realizations of the data were available, converges to the actual value that we want to compare with the model prediction:

$$\langle \hat{D} \rangle = D(\alpha). \tag{7}$$

A measure for its deviations from sample to sample, or the uncertainty in the actual measurement, is then given by the variance of $\hat{D}$, defined as

$$\mathrm{Var}(\hat{D}) = \langle \hat{D}^2 \rangle - \langle \hat{D} \rangle^2. \tag{8}$$

In such a situation, a major role is played by the so-called Cramér–Rao inequality (Rao 1973), which links the Fisher information content of the data to the variance of the estimator, stating that

$$\mathrm{Var}(\hat{D}) F^X(\alpha) \geq \left(\frac{\partial D(\alpha)}{\partial \alpha} \right)^2 . \tag{9}$$

This equation holds for any such estimator $\hat{D}$ and any model parameter α. Two different interpretations of this equation are possible as follows.

The first bounds the variance of $\hat{D}$ by the inverse of the Fisher information. To see this, we consider the special case of the model parameter α being D itself. Although we are making in general a conceptual distinction between the observable D and the model

parameter α, nothing requires us to do so. Since α is now equal to D, the derivative on the right-hand side becomes unity, and one obtains

$$\mathrm{Var}(\hat{D}) \geq \frac{1}{F^X(D)}. \tag{10}$$

The variance of any unbiased estimator $\hat{D}$ of D is therefore bounded by the inverse of the amount of information $F^X(D)$ the data possess on D. If $F^X(D)$ is known it gives a useful lower limit on the error bars that the analysis of the data can put on this observable.

The second reading of the Cramér–Rao inequality, closer in spirit to the present work, is to look at how information is lost by constructing the observable D, and discarding the rest of the data set. For this, we rewrite trivially equation (9) as

$$F^X(\alpha) \geq \left(\frac{\partial D}{\partial \alpha}\right)^2 \frac{1}{\mathrm{Var}(\hat{D})}. \tag{11}$$

The expression on the right-hand side is the ratio of the sensitivity of the observable to the model parameter $(\frac{\partial D}{\partial \alpha})^2$, to the accuracy with which the observable can be extracted from the data, $\mathrm{Var}(\hat{D})$. One of the conceivable approaches in order to estimate the true value of the parameter α is to perform a χ^2 fit to the measured value of D. It is simple to show that this ratio, evaluated at the best-fitting value, is in fact proportional to the expected value of the curvature of $\chi^2(\alpha)$ at this value. Since the curvature of the χ^2 surface describes how fast the value of the χ^2 is increasing when moving away from the best-fitting value, its inverse can be interpreted as an approximation to the error bar that the analysis with the help of $\hat{D}$ will put on α.

Thus, equation (11) shows that by considering only D and not the full data set, we may have lost information on α, a loss given by the difference between the left- and right-hand side of that equation. While the latter may be interpreted as the information on α contained in the part of the data represented by D, we may have lost trace of any other source of information.

2.2 The general case

These considerations on the Cramér–Rao bound can be easily generalized to the case of many parameters and many estimators of as many observables. Dealing with a measurement X with outcomes x, we intend to estimate a set of parameters

$$\boldsymbol{\theta} = (\alpha, \beta, \ldots) \tag{12}$$

with the help of some vector of observables,

$$\boldsymbol{D} = (D_1, \ldots, D_n), \tag{13}$$

which are extracted from x with the help of an array of unbiased estimators,

$$\hat{\boldsymbol{D}} = (\hat{D}_1, \ldots, \hat{D}_n), \quad \langle \hat{\boldsymbol{D}} \rangle = \boldsymbol{D}. \tag{14}$$

In this multidimensional setting, all the three scalar quantities that played a role in our discussion in Section 2.1, i.e. the variance of the estimator, the derivative of the observable with respect to the parameter, and the Fisher information, are now matrices.

The Fisher information $\mathbf{F}$ in X on the parameters $\boldsymbol{\theta}$ is defined as the square matrix:

$$[\mathbf{F}^X(\boldsymbol{\theta})]_{\alpha\beta} = \left\langle \frac{\partial \ln p_X}{\partial \alpha} \frac{\partial \ln p_X}{\partial \beta} \right\rangle. \tag{15}$$

While the diagonal elements are identical to the information scalars in equation (1), the off-diagonal ones describe correlated information. The Fisher information matrix still has the three properties we discussed in Section 2.

The variance of the estimator in equation (8) now becomes the covariance matrix $\mathrm{cov}(\hat{\boldsymbol{D}})$ of the estimators $\hat{\boldsymbol{D}}$, defined as

$$\mathrm{cov}(\hat{\boldsymbol{D}})_{ij} = \langle \hat{D}_i \hat{D}_j \rangle - D_i D_j. \tag{16}$$

Finally, the derivative of the observable with respect to the parameter, on the right-hand side of (9), becomes a matrix $\boldsymbol{\Delta}$, in general rectangular, and defined as

$$\Delta_{\alpha i} = \frac{\partial D_i}{\partial \alpha}, \tag{17}$$

where α runs over all elements of the set $\boldsymbol{\theta}$ of model parameters. Again, the Cramér–Rao inequality provides a useful link between these three matrices, and again there are two approaches to that equation – first, as usually presented in the literature (Rao 1973), in the form of a lower bound to the covariance matrix of the estimators:

$$\mathrm{cov}(\hat{\boldsymbol{D}}) \geq \boldsymbol{\Delta}^{\mathrm{T}} [\mathbf{F}^X(\boldsymbol{\theta})]^{-1} \boldsymbol{\Delta}. \tag{18}$$

The inequality between two symmetric matrices $\mathbf{A} \geq \mathbf{B}$ having the meaning that the matrix $\mathbf{A} - \mathbf{B}$ is positive definite.[1] If, as above, we consider the special case of identifying the parameters with the observables themselves, the matrix $\boldsymbol{\Delta}$ is the identity matrix, and so we obtain that the covariance of the vector of the estimators is bounded by the inverse of the amount of Fisher information on the observables in the data:

$$\mathrm{cov}(\hat{\boldsymbol{D}}) \geq [\mathbf{F}^X(\boldsymbol{D})]^{-1}. \tag{19}$$

Secondly, we can turn this lower bound on the covariance into a lower bound on the amount of information in the data set as well. By rearranging equation (18), we obtain the multidimensional analogue of equation (11), which describes the loss of information that occurs when the data are reduced to a set of estimators:

$$F^X(\boldsymbol{\theta}) \geq \boldsymbol{\Delta} [\mathrm{cov}(\hat{\boldsymbol{D}})]^{-1} \boldsymbol{\Delta}^T. \tag{20}$$

For the sake of completeness, a proof of these two inequalities can be found in Appendix A.

Instead of giving a useful lower bound to the covariance of the estimator as in equation (18), in this form the Cramér–Rao inequality makes it clear how information is in general lost when reducing the data to any particular set of estimators. The right-hand side may be seen, as before, as the expected curvature of a χ^2 fit to the estimates produced by the estimators $\hat{\boldsymbol{D}}$, when evaluated at the best-fitting value, with all correlations fully and consistently taken into account.

In the next two sections, we show how Jaynes' Maximum Entropy Principle allows us to understand the total information content of a data set, once a model is specified, in very similar terms.

2.3 Jaynes' Maximum Entropy Principle

In cosmology, the knowledge of the probability distribution of the data as a function of the parameters, $p_X(x, \boldsymbol{\theta})$, which is compulsory in order to evaluate its Fisher information content, is usually very limited. In a galaxy survey, a data outcome x would be typically the full set of angular positions of the galaxies, together with some redshift estimation if available, to which we may add any other kind

[1] A matrix $\mathbf{A}$ is called positive definite when for any vector $\boldsymbol{x}$ it holds that $\boldsymbol{x}^{\mathrm{T}} \mathbf{A} \boldsymbol{x} \geq 0$. A concrete implication for our purposes is, e.g. that the diagonal entries of the left-hand side of (18) or (19), which are the individual variances of each estimator $\hat{D}_i$, are greater than those of the right-hand side. For many more properties of positive definite matrices, see for instance Bhatia (2007).

of information, such as luminosities, shapes, etc. Our ignorance both of the initial conditions and of many relevant physical processes does not allow us to predict either the galaxy positions in the sky, or all the interconnections with all this additional information. Our predictions of the shape of p_X is thus limited to some statistical properties that are sensitive to the model parameters $\boldsymbol{\theta}$, such as the mean density over some large volume, or certain types of correlation functions.

In fact, even if it were possible to devise some procedure in order to get the exact form of p_X, it may eventually turn out to be useless, or even undesirable, to do so. The incredibly large number of degrees of freedom of such a function is very likely to overwhelm the analyst with a mass of irrelevant details, which may have no relevant significance of their own, or improve the analysis in any meaningful way.

These arguments call for a kind a thermodynamical approach, which would try and capture those aspects of the data which are relevant to our purposes, reducing the number of degrees of freedom in a drastic way. Such an approach already exists in the field of probability theory (Jaynes 1957). It is based on Shannon's concept of entropy of a probability distribution (Shannon 1948).

As we have just argued, our predictive knowledge of $p_X(x, \boldsymbol{\theta})$ is limited to some statistical properties. Let us formalize this mathematically, in a similar way as in Section 2.2. Astrophysical theory gives us a set of constraints on the shape of p_X, in the form of averages of some functions o_i,

$$O_i(\boldsymbol{\theta}) = \langle o_i(x) \rangle(\boldsymbol{\theta}), \quad i = 1, \ldots, n, \tag{21}$$

where p_X enters through the angle brackets. As an example, suppose the data outcome x is a map of the matter density field as a function of position. In this case, one of these constraints O_i could be the mean of the field or its power spectrum, as given by some cosmological model.

The role of this array $\boldsymbol{O} = (O_1, \ldots, O_n)$ is to represent faithfully the physical understanding we have of p_X, according to the model, as a function of the model parameters $\boldsymbol{\theta}$. In the ideal case, some way can be devised to extract each one of these quantities O_i from the data and to confront them to theory.

The set of observables $\boldsymbol{D}$, which we used in Section 2.2, would be a subset of these predictions $\boldsymbol{O}$, and we henceforth refer to $\boldsymbol{O}$ as the 'constraints'.

2.4 Maximum entropy distributions

Although p_X must satisfy the constraints (21), there may still be a very large number of different distributions compatible with these. However, the one that maximizes the value of Shannon's entropy,[2] defined as

$$S = -\int \mathrm{d}x \; p_X(x, \boldsymbol{\theta}) \ln p_X(x, \boldsymbol{\theta}). \tag{22}$$

has a very special status among these distributions.

First introduced by Shannon (Shannon 1948) as a measure of the uncertainty in a distribution on the actual outcome, Shannon's entropy is now the cornerstone of information theory. Jaynes' Maximum Entropy Principle states that the p_X for which this measure S is maximum is the one that best deals with our insufficient knowledge of the distribution, and should be therefore preferred. We refer the reader to Jaynes' work (Jaynes 1983; Jaynes & Bretthorst 2003) and to Caticha (2008) for detailed discussions of the role of entropy in probability theory and for the conceptual basis of maximum entropy methods. Astronomical applications related to some extent to Jaynes' ideas include image reconstruction from noisy data (see e.g. Skilling & Bryan 1984; Starck & Pantin 1996; Maisinger, Hobson & Lasenby 2004, and references therein), mass profiles reconstruction from shear estimates (Bridle et al. 1998; Marshall et al. 2002), as well as model comparison when very small amount of data are available (Zunckel & Trotta 2007). We will see that for our purposes as well it provides us a powerful tool, and that the Maximum Entropy Principle is the ideal complement to Fisher information, fitting very well within our discussions in Section 2 on the Cramér–Rao inequality.

Intuitively, the entropy S of p_X tells us how sharply constrained the possible outcomes x are, and Jaynes' Maximum Entropy Principle selects the p_X which is as wide as possible, but at the same time consistent with the constraints (21) that we put on it. The actual maximum value attained by the entropy S, among all the possible distributions that satisfy (21), is a function of the constraints $\boldsymbol{O}$, which we denote by

$$S(O_1, \ldots, O_n). \tag{23}$$

Of course it is a function of the model parameters $\boldsymbol{\theta}$ as well, since they enter the constraints. As we will see, the shape of that surface as a function of $\boldsymbol{O}$, and thus implicitly as a function of $\boldsymbol{\theta}$, is the key point in understanding the Fisher information content of the data. In the following, in order to keep the notation simple, we will omit the dependence on $\boldsymbol{\theta}$ of most of our expressions, though it will always be implicit.

The problem of finding the distribution p_X that maximizes the entropy (22), while satisfying the set of constraints (21), is an optimization exercise. We can quote the end result (Jaynes 1983, chapter 11; Caticha 2008, chapter 4): the probability density function p_X, when it exists, has the following exponential form:

$$p_X(x) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{n} \lambda_i o_i(x) \right), \tag{24}$$

in which to each constraint O_i is associated a conjugate quantity λ_i, which arises formally as a Lagrange multiplier in this optimization problem with constraints. The conjugate variables λ are also called 'potentials', a terminology that we will use in the following. We will see below in equation (28) that the potentials have a clear interpretation, in the sense that each potential λ_i quantifies how sensitive is the entropy function S in (23) to its associated constraint O_i. The quantity Z, which plays the role of the normalization factor, is called the partition function. Since equation (24) must integrate to unity, the explicit form of the partition function is

$$Z(\lambda_1, \ldots, \lambda_n) = \int \mathrm{d}x \; \exp\left(-\sum_{i=1}^{n} \lambda_i o_i(x) \right). \tag{25}$$

The actual values of the potentials are set by the constraints (21). They reduce, namely, in terms of the partition function, to a system of equations so as to determine the potentials:

$$O_i = -\frac{\partial}{\partial \lambda_i} \ln Z, \quad i = 1, \ldots, n. \tag{26}$$

[2] Formally, for continuous distributions the reference to another distribution is needed to render S invariant with respect to invertible transformations, leading to the concept of the entropy of p_X relative to another distribution q_X, $S = \int \mathrm{d}x \; p_X(x) \ln \frac{p_X(x)}{q_X(x)}$, also called Kullback–Leibler divergence. The quantity defined in the text is more precisely the entropy of $p_X(x)$ relative to a uniform probability density function. For a recent account on this, close in spirit to this work, see Caticha (2008).

The partition function Z is closely related to the entropy S of p_X. It is simple to show that the following relation holds:

$$S = \ln Z + \sum_{i=1}^{n} \lambda_i O_i, \tag{27}$$

and that the values of the potentials can be explicitly written as a function of the entropy, in a relation mirroring equation (26):

$$\lambda_i = \frac{\partial S}{\partial O_i}, \quad i = 1, \ldots, n. \tag{28}$$

Given the nomenclature, it is no surprise that a deep analogy between this formalism and statistical physics exists. Just as the entropy, or partition function, of a physical system determines the physics of the system, the statistical properties of these maximal entropy distributions follow from the functional form of the Shannon entropy or its partition function as a function of the constraints. For instance, the covariance matrix of the constraints is given by

$$\langle [o_i(x) - O_i][o_j(x) - O_j] \rangle = \frac{\partial^2 \ln Z}{\partial \lambda_i \partial \lambda_j}. \tag{29}$$

In statistical physics the constraints can be the mean energy, the volume or the mean particle number, with potentials being the temperature, the pressure and the chemical potential. We refer to Jaynes (1957) for the connection to the physical concept of entropy in thermodynamics and statistical physics.

2.5 The structure of the information in large data sets

With our choice of probabilities p_X given by equation (24), the amount of Fisher information on the parameters $\boldsymbol{\theta} = (\alpha, \beta, \ldots)$ of the model can be evaluated in a straightforward way. The dependence on the model goes through the constraints, or, equivalently, through their associated potentials. It holds therefore that

$$\begin{aligned} \frac{\partial \ln p_X(x)}{\partial \alpha} &= -\frac{\partial \ln Z}{\partial \alpha} - \sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \alpha} o_i(x) \\ &= \sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \alpha} [O_i - o_i(x)], \end{aligned} \tag{30}$$

where the second line follows from the first after application of the chain rule and equation (26). Using the covariance matrix of the constraints given in (29), the Fisher information matrix, defined in (15), can then be written as a double sum over the potentials:

$$F^X_{\alpha\beta} = \sum_{i,j=1}^{n} \frac{\partial \lambda_i}{\partial \alpha} \frac{\partial^2 \ln Z}{\partial \lambda_i \partial \lambda_j} \frac{\partial \lambda_j}{\partial \beta}. \tag{31}$$

There are several ways to rewrite this expression as a function of the constraints and/or their potentials. First, it can be written as a single sum by using equation (26) as

$$F^X_{\alpha\beta} = -\sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \alpha} \frac{\partial O_i}{\partial \beta}. \tag{32}$$

Alternatively, since we will be more interested in using the constraints as the main variables, and not the potentials, we can show using equation (28) that it also takes the form[3]

$$F^X_{\alpha\beta} = -\sum_{i,j=1}^{n} \frac{\partial O_i}{\partial \alpha} \frac{\partial^2 S}{\partial O_i O_j} \frac{\partial O_j}{\partial \beta}. \tag{33}$$

We will use both of these last expressions in the following parts of this work.

Equation (33) presents the total amount of information on the model parameters $\boldsymbol{\theta}$ in the data X, when the model predicts the set of constraints O_i. The amount of information is in the form of a sum of the information contained in each constraint, with correlations taken into account, as on the right-hand side of equation (20). In particular, it is a property of the maximum entropy distributions that if the constraints O_i are not redundant, then it follows that the curvature matrix of the entropy surface $-\partial^2 S$ is invertible and is the inverse of the covariance matrix $\partial^2 \ln Z$ between the observables. To see this explicitly, consider the derivative of equation (26) with respect to the potentials:

$$-\frac{\partial O_i}{\partial \lambda_j} = \frac{\partial^2 \ln Z}{\partial \lambda_i \partial \lambda_j}. \tag{34}$$

The inverse of the matrix on the left-hand side, if it can be inverted, is $-\frac{\partial \lambda_i}{\partial O_j}$, which can be obtained taking the derivative of equation (28), with the result

$$-\frac{\partial \lambda_i}{\partial O_j} = -\frac{\partial^2 S}{\partial O_i \partial O_j}. \tag{35}$$

We have thus obtained in equation (33), combining Jaynes' Maximum Entropy Principle with Fisher's information, the exact expression of the Cramér–Rao inequality (20) for our full set of constraints, but with an equality sign.

We see that the choice of maximum entropy probabilities is fair, in the sense that all the Fisher information comes from what was forced upon the probability density function, i.e. the constraints. No additional Fisher information is added when these probabilities are chosen. In fact, this requirement alone is enough to single out the maximum entropy distributions as being precisely those for which the Cramér–Rao inequality is an equality. This can be understood in terms of sufficient statistics and it goes back to Pitman & Wishart (1936) and Kopman (1936). This was shown in Zografos & Ferentinos (1994). We provide in Appendix A for completeness a similar argument that if the equality sign holds in equation (20) for some distribution, then this is the one that maximizes the entropy relative to some other distribution.

In the special case that the constraints themselves are the model parameters, we have

$$F^X_{O_i O_j} = -\frac{\partial^2 S}{\partial O_i O_j} = -\frac{\partial \lambda_i}{\partial O_j}, \tag{36}$$

which means that the Fisher information on the model predictions contained in the expected future data is given directly by the sensitivity of their corresponding potential. Also, the application of the Cramér–Rao inequality, in the form given in equation (19), to any set of unbiased estimators of $\boldsymbol{O}$ shows that the best joint, unbiased, reconstruction of $\boldsymbol{O}$ is given by the inverse curvature of the entropy surface $-\partial^2 S$, which is, as we have shown, $\partial^2 \ln Z$.

We emphasize at this point that although the amount of information is seen to be identical to the Fisher information in a Gaussian distribution of the observables with the above correlations, nowhere in our approach do we assume Gaussian properties. The distribution of the constraints $o_i(x)$ themselves is set by the maximum entropy distribution of the data.

[3] We note that this result is valid only for maximum entropy distributions and is not equivalent to the second derivative of the entropy with respect to the parameters themselves. However, it is formally identical to the corresponding expression for the information content of distributions within the exponential family (Jennrich & Moore 1975 or van den Bos 2007, chapter 4), once the curvature of the entropy surface is identified with the generalized inverse of the covariance matrix.

2.6 Redundant observables

We have just seen that in the case of independent constraints, the entropy of p_X provides through equation (33) both the joint information content of the data and the inverse correlation matrix between the observables. However, if the constraints put on the distribution are redundant, the correlation matrix is not invertible, and the curvature of the entropy surface cannot be inverted either. We show however that in these cases, our equations for the Fisher information content (31, 32, 33) are still fully consistent, dealing automatically with redundant information to provide the correct answer.

An example of redundant information occurs trivially if one of the functions $o_i(x)$ can be written in terms of the others. For instance, for galaxy survey data, the specification of the galaxy power spectrum as a constraint, together with the mean number of galaxy pairs as a function of distance, and/or the two-point correlation function, which are three equivalent descriptions of the same statistical property of the data. Although the number of observables $\boldsymbol{O}$, and thus the number of potentials, describing the maximum entropy distribution greatly increases by doing so, it is clear that we should expect the Fisher matrix to remain unchanged on addition of such superfluous pieces of information. A small calculation shows that the potentials adjust themselves so that it is actually the case, meaning that this type of redundant information is automatically discarded within this approach. Therefore, we need not worry about the independence of the constraints when evaluating the information content of the data, which will prove convenient in some cases.

There is another, more relevant type of redundant information that allows us to understand better the role of the potentials. Consider that we have some set of constraints $\{O_i\}_{i=1}^{n}$, and that we obtain the corresponding p_X that maximizes the entropy. This p_X could then be used to predict the value O_{n+1} of the average of some other function $o_{n+1}(x)$, which is not contained in our set of predictions:

$$\langle o_{n+1}(x)\rangle =: O_{n+1}. \tag{37}$$

For instance, the maximum entropy distribution built with constraints on the first n moments of p_X will predict some particular value for the $n+1$-th moment, O_{n+1}, that the model was unable to predict by itself.

Suppose now that some new theoretical work provides the shape of O_{n+1} as a function of the model parameters. This new constraint can thus now be added to the previous set, and a new, updated p_X is obtained by maximizing the entropy. There are two possibilities at this point as follows.

The value of O_{n+1} as provided by the model may be identical to the prediction by the maximum entropy distribution that was built without that constraint. Since the new constraint was automatically satisfied, the maximum entropy distribution satisfying the full set of $n+1$ constraints must be equal to the one satisfying the original set. From the equality of the two distributions, which are both of the form (24), it follows that the additional constraint must have a vanishing associated potential,

$$\lambda_{n+1} = 0, \tag{38}$$

while the other potentials are pairwise identical. It follows immediately that the total information as seen from equation (32) is unaffected, and no information on the model parameters was gained by this additional prediction. A cosmological example would be to enforce on the distribution of some field, together with the two-point correlation function, fully disconnected higher order correlation functions. It is well known that the maximum entropy distribution with a constrained two-point correlation function has a Gaussian shape, and that Gaussian distributions have disconnected points function at any order. No information is thus provided by these field moments of higher order in this case.

This argument shows that, for a given set of original constraints and associated maximum entropy distribution, any function $f(x)$, which was not contained in this set, with average F, can be seen as being set to zero potential. Such observables F do not contribute to the information.

More interesting is, of course, the case where this additional constraint differs from the predictions obtained from the original set $\{O_i\}_{i=1}^{n}$. Suppose that there is a mismatch δO_{n+1} between the predictions of the maximum entropy distribution and the model. In this case, when updating p_X to include this constraint, the potentials are changed by this new information, a change given to first order by

$$\delta\lambda_i = \frac{\partial^2 S}{\partial O_i \partial O_{n+1}}\delta O_{n+1}, \quad i = 1, \ldots, n+1, \tag{39}$$

and the amount of Fisher information changes accordingly.

Of course, although the formulae of this section are valid for any model, it requires numerical work in order to get the partition function and/or the entropy surface in a general situation.

2.7 The entropy and Fisher information content of Gaussian homogeneous fields

In order to close this section, we now obtain the Shannon entropy of a family of fields when only the two-point correlation function is the relevant constraint, which we will use extensively in the next section dealing with our cosmological application. It is easily obtained by a straightforward generalization of the finite-dimensional multivariate case, where the means and covariance matrix of the variables are known. It is well known (Shannon 1948) that the maximum entropy distribution is in this case the multivariate Gaussian distribution. Denoting the constraints on p_X by the matrix $\mathbf{D}$ and the vector $\boldsymbol{\mu}$,

$$\begin{aligned} \boldsymbol{D}_{ij} &= \langle x_i x_j\rangle, \\ \mu_i &= \langle x_i\rangle, \quad i, j = 1, \ldots, N, \end{aligned} \tag{40}$$

the associated potentials are given explicitly by the relations

$$\begin{aligned} \boldsymbol{\lambda} &= \frac{1}{2}\mathbf{C}^{-1}, \\ \boldsymbol{\eta} &= -\mathbf{C}^{-1}\boldsymbol{\mu}, \end{aligned} \tag{41}$$

where $\mathbf{C}$ is the covariance matrix,

$$\mathbf{C} := \mathbf{D} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}. \tag{42}$$

The Shannon entropy is given by, up to some irrelevant additive constant,

$$S(\mathbf{D}, \boldsymbol{\mu}) = \frac{1}{2}\ln\det(\mathbf{D} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}). \tag{43}$$

The fact that about half of the constraints are redundant, due to the symmetry of the $\mathbf{D}$ and $\mathbf{C}$ matrices, is reflected by the fact that the corresponding inverse correlation matrix in equation (33),

$$-\frac{\partial^2 S}{\partial D_{ij}\partial D_{kl}} = -\frac{\partial\lambda_{ij}}{\partial D_{kl}} = \frac{1}{2}C^{-1}_{ik}C^{-1}_{jl}, \tag{44}$$

is not invertible as such if we consider all the entries of the matrix $\mathbf{D}$ as constraints. Of course, this is not the case anymore if only the independent entries of $\mathbf{D}$ form the constraints.

2.7.1 Fields, means and correlations

Using the handy formalism of functional calculus, we can straightforwardly extend the above relations to systems with infinite degrees of freedom, i.e. fields, where means as well as the two-point correlation functions are constrained. A realization of the variable X is now a field, or a family of fields $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$, taking values on some n-dimensional space. The expressions above in the multivariate case all stays valid, with the understanding that operations such as matrix multiplications have to be taken with respect to the discrete indices as well as the continuous ones.

With the two-point correlation function and means,

$$\rho_{ij}(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{y}) \rangle, \quad \bar{\phi}_i(\boldsymbol{x}) = \langle \phi_i(\boldsymbol{x}) \rangle, \tag{45}$$

we still have, up to an unimportant constant,

$$S = \frac{1}{2} \ln \det(\rho - \boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}}). \tag{46}$$

In n-dimensional Euclidean space, within a box of volume V for a family of homogeneous fields, it is simplest to work with the spectral matrices. These are defined as

$$\frac{1}{V} \left\langle \tilde{\phi}_i(\boldsymbol{k}) \tilde{\phi}_j^*(\boldsymbol{k}') \right\rangle = P_{ij}(\boldsymbol{k})\, \delta_{\boldsymbol{k}\boldsymbol{k}'}, \tag{47}$$

where the Fourier transforms of the fields are defined through

$$\tilde{\phi}_i(\boldsymbol{k}) = \int_V \mathrm{d}^n x\ \phi_i(x)\, \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}. \tag{48}$$

It is well known that these matrices provide an equivalent description of the correlations, since they form Fourier pairs with the correlation functions:

$$\rho_{ij}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{V} \sum_{\boldsymbol{k}} P_{ij}(\boldsymbol{k}) \mathrm{e}^{\mathrm{i}\boldsymbol{k}(\boldsymbol{x}-\boldsymbol{y})} = \rho_{ij}(\boldsymbol{x} - \boldsymbol{y}). \tag{49}$$

In this case, the entropy in equation (46) reduces, again discarding irrelevant constants, to an uncorrelated sum over the modes,

$$S = \frac{1}{2} \ln \det \left[\frac{\mathbf{P}(0)}{V} - \bar{\boldsymbol{\phi}}\bar{\boldsymbol{\phi}}^{\mathrm{T}} \right] + \frac{1}{2} \sum_{\boldsymbol{k}} \ln \det \frac{\mathbf{P}(\boldsymbol{k})}{V}, \tag{50}$$

which is the straightforward multidimensional version of equation (39) of Taylor & Watts (2001). A comparison with equation (43) shows the well-known fact that the modes can be seen as Gaussian, uncorrelated and complex variables with correlation matrices being proportional to $\mathbf{P}(\boldsymbol{k})$. All the modes have zero mean, except for the zero-mode, which, as seen from its definition, is proportional to the mean of the field itself. Accordingly, taking the appropriate derivatives, the potentials $\boldsymbol{\lambda}(\boldsymbol{k})$ associated to $\mathbf{P}(\boldsymbol{k})$ read

$$\boldsymbol{\lambda}(\boldsymbol{k}) = \frac{V}{2} \mathbf{P}(\boldsymbol{k})^{-1}, \quad \boldsymbol{k} \neq 0, \qquad \boldsymbol{\lambda}(0) = \frac{1}{2} \left[\frac{\mathbf{P}(0)}{V} - \boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}} \right]^{-1}, \tag{51}$$

and those associated to the mean $\boldsymbol{\phi}$ read

$$\boldsymbol{\eta} = -\left[\frac{\mathbf{P}(0)}{V} - \boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}} \right]^{-1} \boldsymbol{\phi}. \tag{52}$$

Note that although the spectral matrices are, in general, complex, they are hermitian, so that the determinants are real. The amount of Fisher information in the family of fields is easily obtained with the help of equation (32), with the familiar result

$$F_{\alpha\beta} = \frac{1}{2} \sum_{\boldsymbol{k}} \mathrm{Tr} \left[\mathbf{P}_{\mathrm{c}}^{-1}(\boldsymbol{k}) \frac{\partial \mathbf{P}_{\mathrm{c}}(\boldsymbol{k})}{\partial \alpha} \mathbf{P}_{\mathrm{c}}^{-1}(\boldsymbol{k}) \frac{\partial \mathbf{P}_{\mathrm{c}}(\boldsymbol{k})}{\partial \beta} \right] + \frac{\partial \bar{\boldsymbol{\phi}}^{\mathrm{T}}}{\partial \alpha} \left[\frac{\mathbf{P}_{\mathrm{c}}(0)}{V} \right]^{-1} \frac{\partial \bar{\boldsymbol{\phi}}}{\partial \beta}, \tag{53}$$

with $\mathbf{P}_{\mathrm{c}}(\boldsymbol{k})$ being the connected part of the spectral matrices,

$$\mathbf{P}_{\mathrm{c}}(\boldsymbol{k}) = P(\boldsymbol{k}) - \delta_{k0} V \boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}}. \tag{54}$$

These expressions are of course also valid for isotropic fields on the sphere. With a decomposition in spherical harmonics, the sum runs over the multipoles.

3 COSMOLOGICAL APPLICATION TO WEAK LENSING OBSERVABLES

Gravitational lensing, which can be used to measure the distribution of mass along the line of sight, has been recognized as a powerful probe of the dark components of the Universe (Schneider et al. 1992; Bartelmann & Schneider 2001; Refregier 2003; Munshi et al. 2006; Schneider et al. 2006) since it is sensitive both to the geometry of the Universe and to the growth of structure. Weak lensing data are typically used in two ways. The first, which is deployed for cosmological parameter fitting, relies on measuring the correlated distortions in galaxy images (Albrecht et al. 2006). The second approach uses each galaxy to make a noisy measurement of the lensing signal at that position. These point estimates are then used to reconstruct the dark matter density distribution (e.g. Kaiser & Squires 1993; Seitz & Schneider 2001). Most of the measurements of weak lensing to date have focused on the shearing that galaxy images experience. However, gravitational lensing causes a number of other distortions of galaxy images. These include change in size, which is related to the magnification, and higher order image distortions known as flexion (Bacon et al. 2006). A number of techniques have been developed for measuring these higher order image distortions, such as the HOLICS (Okura, Umetsu & Futamase 2007) and shapelets methods (Massey et al. 2007). Since all of the image distortions originate in the same cause, i.e. the lensing potential field, the information content of any two lensing measurements must be degenerate. At the same time, since each method has different systematics and specific noise properties, combining multiple measurements may bring substantial benefits. Some recent works have looked at the impact of combining shear and flexion measurements for mass reconstruction (Er, Li & Schneider 2010; Pires & Amara 2010; Velander, Kuijken & Schrabback 2011) as well as the benefits for breaking multiplicative bias of including galaxy size measurements (Vallinotto, Dodelson & Zhang 2010).

3.1 Linear probes

The predictive power of some observable O_{c} of a central field (for instance its power spectrum at some mode) translates into an array of constraints O_i, $i = 1, \ldots, n$, in the noisy probes, that we could try and extract and confront with theory:

$$O_i(\boldsymbol{\theta}) = f_i(O_{\mathrm{c}}(\boldsymbol{\theta})), \quad i = 1, \ldots, n \tag{55}$$

for some functions f_i.

For the purpose of this work, the case of functions linear with respect to O_{c} is generic enough, i.e. we will consider that

$$\frac{\partial^2 f_i}{\partial O_{\mathrm{c}}^2} = 0, \quad i = 1, \ldots, n. \tag{56}$$

The entropy S of the data is a function of the n constraints $\boldsymbol{O}$. It is, however, fundamentally a function of O_c since it does enter all of these observables. It is therefore very natural to associate a potential λ_c to O_c, although it is not itself a constraint on the probability density function. In analogy with

$$\lambda_i = \frac{\partial S}{\partial O_i}, \quad i = 1, \ldots, n, \tag{57}$$

we define

$$\lambda_c := \frac{\mathrm{d}S}{\mathrm{d}O_c} \quad (O_1, \ldots, O_m) \tag{58}$$

with the result, given by an application of the chain rule, of

$$\lambda_c = \boldsymbol{\lambda} \cdot \frac{\partial \boldsymbol{f}}{\partial O_c}. \tag{59}$$

On the other hand, the impact of a model parameter on each observable can be similarly written in terms of the central observable O_c:

$$\frac{\partial \boldsymbol{O}}{\partial \alpha} = \frac{\partial O_c}{\partial \alpha} \frac{\partial \boldsymbol{f}}{\partial O_c}. \tag{60}$$

It follows directly from relations (59) and (60), and from the linearity of f_i, that the joint information in the full set of constraints $\boldsymbol{O}$, given in equation (32) as a sum over all the n constraints, reduces to a formally identical expression with the difference that O_c only does enter,

$$F^X_{\alpha\beta} = \frac{\partial \boldsymbol{O}}{\partial \alpha} \cdot \frac{\partial \boldsymbol{\lambda}}{\partial \beta} = \frac{\partial \lambda_c}{\partial \alpha} \frac{\partial O_c}{\partial \beta}, \tag{61}$$

which can also be written in a form analogous to (33):

$$F^X_{\alpha\beta} = -\frac{\partial O_c}{\partial \alpha} \frac{\mathrm{d}^2 S}{\mathrm{d}O_c^2} \frac{\partial O_c}{\partial \beta}. \tag{62}$$

This last equation shows that all the effects of combining this set of constraints have been absorbed into the second total derivative of the entropy. This second total derivative is the total amount of information there is on the central quantity O_c in the data. Indeed, taking as a special case of model parameter to the central quantity itself, i.e.

$$\alpha = \beta = O_c, \tag{63}$$

one obtains now that the full amount of information in X on O_c is

$$F^X_{O_c O_c} = -\frac{\mathrm{d}^2 S}{\mathrm{d}O_c^2}(O_1, \ldots, O_n) \equiv \frac{1}{\sigma^2_{\mathrm{eff}}}. \tag{64}$$

A simple application of the Cramer–Rao inequality presented in equation (11) shows that this effective variance σ_{eff} is the lower bound to an unbiased reconstruction of the central observable from the noisy probes.

These considerations on the effect of probe combination in the case of a single central field observable O_c generalize easily to the case where there are many $(O_c^1, \ldots, O_c{}^m)$. In this case, each central field quantity leads to an array of constraints in the form of equation (55), therefore it is simple to show that the amount of Fisher information can again be written in terms of the information associated to the central field, with an effective covariance matrix between the O_c values. The result is

$$F^X_{\alpha\beta} = -\sum_{i,j=1}^{m} \frac{\partial O_c^i}{\partial \alpha} \frac{\mathrm{d}^2 S}{\mathrm{d}O_c^i O_c^j} \frac{\partial O_c^j}{\partial \beta}. \tag{65}$$

All the effects of probe combination are thus encompassed in an effective covariance matrix $\boldsymbol{\Sigma}_{\mathrm{eff}}$ of the central field observables:

$$-\frac{\mathrm{d}^2 S}{\mathrm{d}O_c^i O_c^j} \equiv \left[\boldsymbol{\Sigma}^{-1}_{\mathrm{eff}}\right]_{ij}. \tag{66}$$

Again, an application of the Cramer–Rao inequality in the multi-dimensional case shows that this effective covariance matrix is the best achievable unbiased joint reconstruction of $(O_c^1, \ldots, O_c^m)$.

We now explore further the case of linear probes of homogeneous Gaussian fields, which is cosmologically relevant and can be solved analytically to full extent. We will focus on zero mean fields, for which, according to our previous section, the entropy can be written in terms of the spectral matrices, up to a constant:

$$S = \frac{1}{2} \sum_{\boldsymbol{k}} \ln \det \mathbf{P}(\boldsymbol{k}). \tag{67}$$

3.2 Linear tracers at the two-point level

A standard instance of a linear tracer ϕ_i of some central field κ in weak lensing is provided by a relation in Fourier space of the form of

$$\tilde{\phi}_i(\boldsymbol{k}) = v_i \tilde{\kappa}(\boldsymbol{k}) + \tilde{\epsilon}_i(\boldsymbol{k}) \tag{68}$$

for some noise term $\tilde{\epsilon}_i$, uncorrelated with κ, and coefficient v_i. Typically, if one observes a tracer of the derivative of the field κ, then the vector $\boldsymbol{v}$ would be proportional to $-\mathrm{i}\boldsymbol{k}$. We are ignoring here any observational effect, such as incomplete sky coverage, that would require corrections to this relation. It is clear from this relation that the spectral matrices of this family take the special form of equation (55): defining the spectrum of the κ field by P^κ, we obtain by putting this relation (68) into (47), that the spectral matrices can be written in each mode in the form of

$$\boldsymbol{P} = P^\kappa \boldsymbol{v}\boldsymbol{v}^\dagger + \mathbf{N}, \tag{69}$$

where $\boldsymbol{v}^\dagger$ is the hermitian conjugate of $\boldsymbol{v} = (v_1, \ldots, v_n)$. The matrix $\mathbf{N}$ is the spectrum of the noise components ϵ:

$$N_{ij}(\boldsymbol{k}) = \frac{1}{V} \langle \tilde{\epsilon}_i(\boldsymbol{k}) \tilde{\epsilon}^*(\boldsymbol{k}) \rangle. \tag{70}$$

Our subsequent results hold for any family of tracers that obey this relation. While the special case of (68) falls in this category, this need not be the only instance. All the weak lensing observables we deal with in this work will satisfy equation (69).

Both the n-dimensional vector $\boldsymbol{v}$ and the noise matrix $\mathbf{N}$ can depend on the wave vector $\boldsymbol{k}$, but they are independent of the model parameters. The matrix $\mathbf{N}$ of dimension $n \times n$ is the noise component of the spectra of the fields, typically built from two parts. The first is due to the discrete nature of the fields, since such data consist of quantities measured where galaxies sit, and the second to the intrinsic dispersion of the measured values.

3.3 Joint entropy and information content

Information on the model parameters enters through P^κ only. To evaluate the full information content, we need only to evaluate equation (67) with the spectral matrix given in (69), keeping in mind the result from the last section that we need only the total derivative with respect to P^κ. In other words, any additive terms in the expression of the entropy that are independent of P^κ can be discarded.

This determinant can be evaluated immediately. Defining for each mode the real positive number N_{eff} through

$$\frac{1}{N_{\mathrm{eff}}} \equiv \boldsymbol{v}^\dagger \boldsymbol{N}^{-1} \boldsymbol{v}, \tag{71}$$

which can be seen as an effective noise term, a simple[4] calculation shows that the joint entropy (67) is equivalent to the following, where the n-dimensional determinant has disappeared:

$$S = \frac{1}{2}\sum_{\boldsymbol{k}} \ln[P^{\kappa}(\boldsymbol{k}) + N_{\rm eff}(\boldsymbol{k})]. \tag{73}$$

Comparison with equation (67) shows that we have with equation (73) the entropy of the field κ itself, where all the effects of the joint observation of these n fields have been absorbed into the effective noise term $N_{\rm eff}$, which contaminates its spectrum. It means that the full combined information in the n probes of the field κ is equivalent to the information in κ, observed with spectral noise $N_{\rm eff}$.

Our result (64) applied to (73) puts bounds on the reconstruction of the field κ out of the observed samples, which can be at best reconstructed with a contaminating noise term of $N_{\rm eff}$ in its spectrum, whose best unbiased reconstruction is given by

$$2[P^{\kappa}(\boldsymbol{k}) + N_{\rm eff}(\boldsymbol{k})]^2. \tag{74}$$

Since the effect of combining these probes at a single mode is only to change the model-independent noise term, the parameter correlations and degeneracies as approximated by the Fisher information matrix stay unchanged, whatever the number of such probes is. We have namely from (73) that at a given mode $\boldsymbol{k}$, the Fisher information matrix reads

$$F^{X}_{\alpha\beta} = \frac{1}{2}\frac{\partial \ln \tilde{P}^{\kappa}(\boldsymbol{k})}{\partial\alpha}\frac{\partial \ln \tilde{P}^{\kappa}(\boldsymbol{k})}{\partial\beta} \tag{75}$$

with

$$\tilde{P}^{\kappa}(\boldsymbol{k}) = P^{\kappa}(\boldsymbol{k}) + N_{\rm eff}(\boldsymbol{k}). \tag{76}$$

From the point of view of the Fisher information, it makes formally no difference to extract the full set of $n(n-1)/2$ independent elements of each spectral matrix, or reconstruct the field κ and extract its spectrum. They carry indeed the same amount of Fisher information.

These results still hold when other fields are present in the analysis, which are correlated with the field κ. To make this statement rigorous, consider in the analysis on top of our n samples of the form (68) of κ, another homogeneous field θ, with spectrum $P^{\theta}(\boldsymbol{k})$, and cross-spectrum to κ given by $P^{\theta\kappa}(\boldsymbol{k})$ The full spectral matrices are in this case

$$\mathbf{P}(\boldsymbol{k}) = \begin{pmatrix} P^{\kappa}(\boldsymbol{k})\boldsymbol{v}\boldsymbol{v}^{\rm T} + \boldsymbol{N} & P(\boldsymbol{k})^{\kappa\theta}\boldsymbol{v} \\ P^{\theta\kappa}\boldsymbol{v}^{\rm T} & P^{\theta}(\boldsymbol{k}) \end{pmatrix}. \tag{77}$$

Again, the determinant of this matrix can be reduced to a determinant of lower dimension, leading to the equivalent entropy

$$S = cst + \frac{1}{2}\ln\det\begin{pmatrix} P^{\psi}(\boldsymbol{k}) + N_{\rm eff} & P^{\kappa\theta}(\boldsymbol{k}) \\ P^{\theta\kappa}(\boldsymbol{k}) & P^{\theta}(\boldsymbol{k}) \end{pmatrix}. \tag{78}$$

It shows that the full set of $n+1$ fields can be reduced without loss to two fields, κ and θ, with the effective noise $N_{\rm eff}$ contaminating the spectrum of κ.

Note that the derivation of our results do not refer to any hypothetical estimators, but came naturally out of the expression of the entropy.

[4] We have namely for any invertible matrix $\mathbf{A}$ and vectors $\boldsymbol{u}$, $\boldsymbol{v}$ the matrix determinant lemma

$$\det(\mathbf{A} + \boldsymbol{u}\boldsymbol{v}^{\rm T}) = \det(\mathbf{A})(1 + \boldsymbol{v}^{\rm T}\mathbf{A}^{-1}\boldsymbol{u}). \tag{72}$$

3.4 Weak lensing probes

We now seek a quantitative evaluation of the full joint information content of the weak lensing probes in galaxy surveys, up to second order in the image distortions of galaxies. The data X consist of a set of fields, which are discrete point fields, which take values where galaxies sit. We work in the two-dimensional flat sky limit, using the more standard notation $\boldsymbol{l}$ for the wavevector, and decompose it in modulus and polar coordinate as

$$\boldsymbol{l} = l\begin{pmatrix}\cos\varphi_l \\ \sin\varphi_l\end{pmatrix}. \tag{79}$$

For the scope of this paper, we will throughout assume that the intrinsic values of each probe are pairwise uncorrelated, as commonly done. Also, we will assume that the set of points on which the relevant quantities are measured show a low enough clustering so that corrections to the spectra due to intrinsic clustering can be ignored. This is, however, not a limitation of our approach, since corrections to the above assumptions, such as the introduction of some level of intrinsic alignment, can be accommodated by introducing appropriate terms in the noise matrices $N(\boldsymbol{k})$ in (71). As a central field to which all our point fields relate, we take for convenience the isotropic convergence field κ, with spectrum

$$C^{\kappa}(\boldsymbol{l}) = C^{\kappa}(l). \tag{80}$$

In the case of pairwise uncorrelated intrinsic values that we are following, we see easily from (71) that by combining any number of such probes the effective noise is reduced at a given mode according to

$$\frac{1}{N^{\rm tot}_{\rm eff}} = \sum_i \frac{1}{N^{i}_{\rm eff}}. \tag{81}$$

We therefore only need to evaluate the effective noise for each probe separately, while their combination follows (81). To this aim, the evaluation of the spectral matrices (69), giving us $N_{\rm eff}$, is necessary. The calculations for this are presented in Appendix B and we use the final results in this section.

3.4.1 First order, distortion matrix

To first order, the distortion induced by weak lensing on a galaxy image is described by the distortion matrix that contains the shear, γ, and convergence, κ, which come from the second derivatives of the lensing potential field ψ (e.g. Schneider et al. 2006):

$$\begin{pmatrix}\kappa+\gamma_1 & \gamma_2 \\ \gamma_2 & \kappa-\gamma_1\end{pmatrix} = \psi_{,ij}. \tag{82}$$

The shear components read

$$\gamma_1 = \frac{1}{2}(\psi_{,11} - \psi_{,22}), \quad \gamma_2 = \psi_{,12}, \tag{83}$$

and we assume that they are measured from the apparent ellipticities of the galaxies, with identical intrinsic dispersion σ^2_{γ}. Denoting by $\bar{n}_{\gamma}$ the number density of galaxies for which ellipticity measurements are available, the effective noise is

$$N^{\gamma}_{\rm eff} = \frac{\sigma^2_{\gamma}}{\bar{n}_{\gamma}}. \tag{84}$$

The information content of the two observed ellipticity fields is thus exactly the same as the one of the convergence field, with a mode-independent noise term as above.

To reach for the κ component of the distortion matrix, we imagine we have measurements of their angular size $s_{\rm obs}$, with intrinsic dispersion $\sigma_{\rm s}^2$. The intrinsic sizes of the galaxies $s_{\rm int}$ gets transformed through weak lensing according to

$$s_{\rm obs} = s_{\rm int}(1 + \alpha_{\rm s}\kappa). \tag{85}$$

The coefficient $\alpha_{\rm s}$ is equal to unity in pure weak lensing theory, but we allow it to take other values since in a realistic situation other effects such as magnification bias effectively enter this coefficient (e.g. Vallinotto et al. 2010). Under our assumption that the correlation between the fluctuations in intrinsic sizes can itself be ignored, the effective noise reduces to

$$N_{\rm eff}^{\rm s} = \frac{1}{\alpha_{\rm s}^2}\left(\frac{\sigma_{\rm s}}{\bar{s}_{\rm int}}\right)^2 \frac{1}{\bar{n}_{\rm s}}. \tag{86}$$

This combination of $\alpha_{\rm s}$ with the dispersion parameters $\bar{s}$ and $\sigma_{\rm s}$ becomes the only relevant parameter in our case, and not the value of each of them.

3.4.2 *Second order, flexion*

To second order, the distortions caused by lensing on the galaxies' images are given by third-order derivatives of the lensing potential. These are conveniently described by the spin 1 and spin 3 flexion components $\mathcal{F}$ and $\mathcal{G}$, which in the notation of Schneider & Er (2008) read

$$\begin{aligned}\mathcal{F} &= \frac{1}{2}\begin{pmatrix}\psi_{,111} + \psi_{,122}\\ \psi_{,112} + \psi_{,222}\end{pmatrix},\\ \mathcal{G} &= \frac{1}{2}\begin{pmatrix}\psi_{,111} - 3\psi_{,122}\\ 3\psi_{,112} - \psi_{,222}\end{pmatrix},\end{aligned} \tag{87}$$

and are extracted from measurements with intrinsic dispersion $\sigma_{\mathcal{F}}^2$ and $\sigma_{\mathcal{G}}^2$. The effective noise this time is mode-dependent:

$$\frac{1}{N_{\rm eff}^{\mathcal{FG}}} = l^2\left(\frac{\bar{n}_F}{\sigma_F^2} + \frac{\bar{n}_G}{\sigma_G^2}\right). \tag{88}$$

4 RESULTS

Fig. 1 shows the ratio of the effective noise to the noise present considering the shear fields only, assuming the same number densities of galaxies for each probe, and the values for the intrinsic dispersion stated in Table 1. The conversion multipole l (upper x-axis) to angular scale θ (lower x-axis) follows $\theta = \pi/(l + 1/2)$. We have adopted for the size dispersion parameters the numbers from Vallinotto et al. (2010), who evaluated this number for the DES survey conditions (The Dark Energy Survey Collaboration 2005). We refer to the discussion in Pires & Amara (2010) for our choice of flexion dispersion parameters. The curves on this figure are ratios and therefore independent of the galaxy number density. They are redshift independent as well, only to the extent that the dispersion in intrinsic values can be treated as such. We can draw two main conclusions from Fig. 1. First, flexion information begins to play a role only at the smallest scales, i.e. on the arcsecond scales, where it takes over and becomes the most interesting probe. On the scale of 1 arcmin, it can bring substantial improvement over shear only analysis, but only in combination with the shears, and not on its own. This is in a good agreement with the comparative analysis of the power of the flexion $\mathcal{F}$ field and shear fields for mass reconstruction done in Pires & Amara (2010), restricted to direct inversion

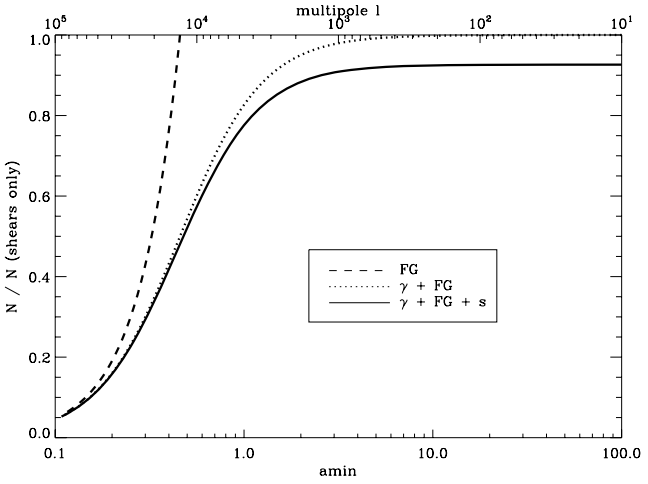

Figure 1. The ratio of the effective noise to the level of noise considering the shears only, as a function of angular scale. The dashed line considers the flexion fields alone. The dotted line shows the combination of the flexion fields with the shear fields, and the solid line is all these weak lensing probes combined. No correlations between the intrinsic values for each pair of probes have been considered.

Table 1. Dispersion parameters used in Fig. 1.

σ_γ	$\sigma_{\mathcal{F}}$ (arcsec^{-1})	$\sigma_{\mathcal{G}}$ (arcsec^{-1})	$\frac{1}{\alpha_{\rm s}}\frac{\sigma_{\rm s}}{\bar{s}}$
0.25	0.04	0.04	0.9

methods. Secondly, the inclusion of size of galaxies into the analysis provides a density-independent, scale-independent improvement factor of

$$\frac{N_{\rm eff}^{\gamma}}{N_{\rm eff}^{\gamma+{\rm s}}} = 1 + \left(\frac{\sigma_\gamma \bar{s}\alpha_{\rm s}}{\sigma_{\rm s}}\right)^2, \tag{89}$$

which is close to a 10 per cent improvement for the quoted numbers. Of course, the precise value depends on the dispersion parameters of the population considered.

For the purpose of measuring cosmological parameters rather than mass reconstruction, more interesting are the actual values of the Fisher information matrices. Since with any combination of such probes, these matrices are proportional to each other in a single mode, it makes sense to define the efficiency parameter of the probe i through

$$\epsilon_i(l) := \frac{C^\kappa(l)}{C^\kappa(l) + N_{\rm eff}^i(l)}, \tag{90}$$

which is a measure of what fraction of the information contained in the convergence field is effectively caught by that probe. The information in the convergence field is, at a given mode l, counting the multiplicity of the mode,

$$F_{\alpha\beta}^\kappa = \frac{1}{2}(2l+1)\frac{\partial \ln C^\kappa(l)}{\partial\alpha}\frac{\partial \ln C^\kappa(l)}{\partial\beta}, \tag{91}$$

and we indeed obtain the total Fisher information in the observed fields:

$$F_{\alpha\beta}^X = \sum_l F_{\alpha\beta}^\kappa(l)\epsilon_i^2(l). \tag{92}$$

Therefore, according to the interpretation of the Fisher matrix approximating the expected constraints on the model parameters, the

factor $\epsilon(l)$ is precisely equal to the factor of degradation in the constraints one would be able to put on any parameter, with respect to the case of a perfect knowledge of the convergence field at this mode. It is not the purpose of this work to perform a very detailed study on the behaviour of the efficiency parameter for some specific survey and the subsequent statistical gain, but its qualitative behaviour is easy to see. This parameter is essentially unity in the high signal-to-noise ratio regime, while it is the inverse effective noise whenever the intrinsic dispersion dominates the observed spectrum. Since information on cosmological parameters is beaten down by cosmic variance in the former case, the latter dominates the constraints. We can therefore expect from our above discussion the size information to tighten by a few per cent of constraints on any cosmological parameter. On the other hand, while flexion becomes ideal for mass reconstruction purposes on small scales, it will be able to help us infer on cosmological parameters only if the challenge of very accurate theoretical predictions on the convergence power spectrum for multipoles substantially larger than 1000 will be met.

To make these expectations more concrete, we evaluated the improvement in information on cosmological parameters performing a lensing Fisher matrix calculation for a wide, *EUCLID*-like survey, in a tomographic setting. For a data vector consisting of n probes of the convergence field κ_i in each redshift bin $i, i = 1, \ldots, N$, it is simple to see following our previous argument that the Fisher information reduces to

$$F_{\alpha\beta} = \frac{1}{2}\sum_l (2l+1)\mathrm{Tr}\, C^{-1}\frac{\partial C}{\partial \alpha}C^{-1}\frac{\partial C}{\partial \beta}, \tag{93}$$

where the $\mathbf{C}$ matrix is given by

$$C_{ij} = C^{\kappa_i\kappa_j}(l) + \delta_{ij}N^i_{\mathrm{eff}}(l), \quad i, j = 1, N \tag{94}$$

with N^i_{eff} given by (71). The only difference between standard implementations of Fisher matrices for lensing, such as the lensing part of Hu & Jain (2004), being thus the form of the noise component. We evaluated these matrices respectively for

$$N^i_{\mathrm{eff}} = \frac{\sigma^2_\gamma}{\bar{n}^i} = N^{\gamma,i}_{\mathrm{eff}}, \tag{95}$$

which is the precise form of the Fisher matrix for shear analysis, for

$$\frac{1}{N^i_{\mathrm{eff}}} = \frac{1}{N^{\gamma,i}_{\mathrm{eff}}} + \frac{1}{N^{s,i}_{\mathrm{eff}}}, \tag{96}$$

which accounts for size information, and

$$\frac{1}{N^i_{\mathrm{eff}}(l)} = \frac{1}{N^{\gamma,i}_{\mathrm{eff}}} + \frac{1}{N^{s,i}_{\mathrm{eff}}} + \frac{1}{N^{\mathcal{FG},i}_{\mathrm{eff}}(l)}, \tag{97}$$

which accounts for the flexion fields as well. We note that in terms of observables, these small modifications incorporate in its entirety the full set of all possible correlations between the fields considered. The values of the dispersion parameters involved in these formulae are the same as in Table 1. Our fiducial model is a flat ΛCDM universe, with parameters $\Omega_\Lambda = 0.7, \Omega_\mathrm{b} = 0.045, \Omega_\mathrm{m} = 0.3, h = 0.7$, power spectrum parameters $\sigma_8 = 0.8$, $n = 1$, and the Chevallier-Polarski-Linder parametrization (Chevallier & Polarski 2001; Linder 2003) of the dark energy equation of state implemented as $\omega_0 = -1$, $w_a = 0$. The distribution of galaxies as a function of redshift needed both for the calculation of the spectra and to obtain the galaxy densities in each bin was generated using the cosmological package ICOSMO (Refregier et al. 2011), in a way described in Amara & Réfrégier (2007). We adopted *EUCLID*-like parameters of 10 redshift bins, a median redshift of 1, a galaxy angular density of 40 arcmin^{-2} and photometric redshift errors of $0.03(1+z)$.

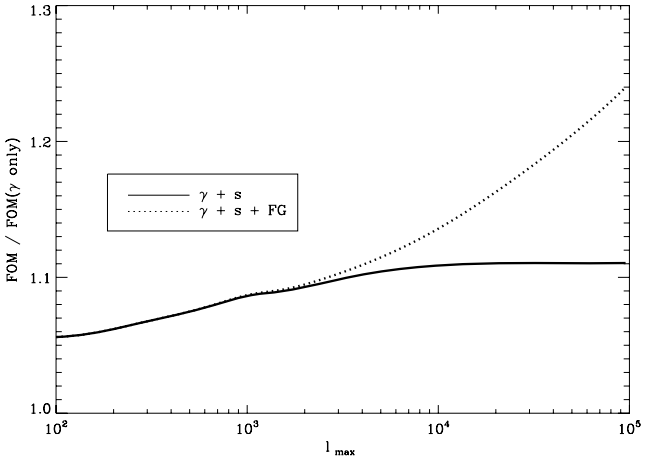

Figure 2. The improvement of the dark energy FOM including size information (solid) as well as flexion $\mathcal{F}$ and $\mathcal{G}$ information (dotted), over the shear-only analysis, as a function of the maximum angular multipole included in the analysis.

Table 2. Ratio of the marginalized constraints $\sigma^2/\sigma^2_{\text{shear only}}$, for $l_{\max} = 10^4$. The first line considers the inclusion of the size information in the analysis, while the second line considers the size as well as the flexion fields $\mathcal{F}$ and $\mathcal{G}$.

Ω_Λ	Ω_b	Ω_m	h	n	σ_8	w_0	w_a
0.90	0.96	0.90	0.95	0.95	0.90	0.90	0.90
0.88	0.96	0.89	0.95	0.93	0.88	0.88	0.88

In Fig. 2, we show the improvement in the dark energy Figure of Merit (FOM), defined as the square root of the determinant of the submatrix (ω_0, ω_a) of the Fisher matrix inverse $\mathbf{F}^{-1}_{\alpha\beta}$ (α and β running over the set of eight parameters as described above), as a function of the maximum angular mode $l_{\max}$ considered, while $l_{\min}$ being always taken to be 10. In perfect agreement with our discussion above, including size information (solid line) increases the FOM steadily until it saturates at a 10 per cent improvement when constraints on the dark energy parameters are dominated by the low signal-to-noise ratio regime. Also, flexion becomes only useful in the deep non-linear regime, where, however, a theoretical understanding of the shape of the spectra still leaves a lot to be desired.

These results are found to be very insensitive to the survey parameters, for a fixed α_s. These are also only weakly model parameter independent, as illustrated in Table 2, which shows the corresponding improvement in Fisher constraints,

$$\frac{\sigma^2}{\sigma^2_{\text{shear only}}} = \frac{F^{-1}_{\alpha\alpha}}{F^{-1}_{\alpha\alpha,\text{shear only}}}, \tag{98}$$

at the saturation scale $l_{\max} = 10^4$. These results are also essentially unchanged using either standard implementations of the halo model (see Cooray & Sheth 2002, for a review) or the HALOFIT (Smith et al. 2003) non-linear power spectrum.

5 SUMMARY AND DISCUSSION

We have shown how Jaynes' Maximum Entropy Principle allows us to construct the Fisher information content on model parameters in a given data set in the form of equation (32) or (33). This is done by making the key quantity the entropy of the distribution as

a function of the constraints that we put on it. These constraints form our knowledge of the statistical properties of the future data. To the best of the authors' knowledge, equation (32) or (33) is not to be found in this form in the literature. However, they cannot be considered new, since as stated earlier, they can be easily derived from the Fisher information content of the exponential family of distributions (Jennrich & Moore 1975; van den Bos 2007), after the identification of the curvature of the entropy surface with the generalized inverse of the covariance matrix. Especially, the maximum entropy distributions are precisely those for which the Cramér–Rao inequality is an equality, since the curvature of the entropy surface is the inverse correlation matrix between the model predictions. Equation (33) also bears a strong formal similarity to the well-known result [Kullback (1959, chap. 2) or Caticha (2008, chap. 6)] that the Fisher information can always be written as the curvature of the Kullback–Leibler divergence for distributions parametrized by the same set of parameters.

The Fisher matrices currently used in weak lensing or clustering can all be seen as special cases of this approach, namely equation (53), when knowledge of the statistical properties of the future data does not go beyond the two-point statistics. Indeed, in the case that the model does not predict the means, and knowing that for discrete fields the spectral matrices, equation (47), carry a noise term due to the finite number of galaxies, or, in the case of weak lensing, also due to the intrinsic ellipticities of galaxies, the amount of information in (53) is essentially identical to the standard expressions used to predict the accuracy with which parameters will be extracted from power-spectra analysis.

There is, however, a conceptual difference worth noting in that the standard approach is to pick an estimator for the power spectra and assume that both the fields and the distribution of the estimators are Gaussian. The result is the amount of Fisher information there is in the power spectra, under the assumptions of Gaussian statistics for the estimators and the fields. In our approach, the only assumption is on the fields' distribution. Our results do not depend on the way the information will be extracted, but shows the amount of Fisher information in the fields as a whole.

Of course, the maximum entropy approach, which tries to capture the relevant properties of p_X through a sophisticated guess, gives no guarantee that its predictions are actually correct. Nevertheless, as discussed in Section 2.6, it provides a systematic approach with which to update the probability density function in case of improved knowledge of the relevant physics.

Using this formalism we have investigated the combined Fisher information content of weak lensing probes up to second order in the shapes distortions, assuming model parameter independent noise. By having a look at the joint Shannon entropy of the fields, we have shown how the only effect of treating these observables jointly is to reduce the effective level of noise contaminating the convergence field, according to equations (71) and (73), independently of the model parameters.

The following are the key points of this paper that we would like to emphasize.

(i) Equation (33) presents a measure of information content that depends only on the constraints put on the data and the physical model. It is written in terms of the curvature of Shannon's entropy surface for maximum entropy distributions. It can always be interpreted, regardless of the actual distribution of the parameters, and of the specific way the analysis will proceed, as the expected curvature of the χ^2 surface to the full set of model predictions. Assumptions of Gaussianity are neither needed nor used at any point.

(ii) Over a very wide range of scales, the probe of choice both for mass reconstruction or cosmological purposes are the ellipticity components of the galaxies. Flexion takes over only on the arcsecond scale. In combination with the ellipticities, it can lead to a substantial increase in statistical power on the scale of arcminute. From the cosmological point of view, we expect size information to contribute at the 10 per cent level of the total information content. The only key parameter is the combination (89) of the dispersion values and the permeability α_s of the population sizes to the convergence field. On the other hand, the prospects of including flexion in cosmological analysis are less clear. The most obvious drawback is the need for an accurate understanding of the non-linear power spectrum.

(iii) Besides, our results render the inclusion of flexion and size information within more detailed Fisher matrix analysis for future dark energy experiments extremely simple, such as in the exhaustive approach combining the information galaxy density fields with shear fields in the tomographic setting of Bernstein (2009). From (78) follows namely that the inclusion of all the two-point correlations of these additional weak lensing probes can be accounted for by adopting the noise term $N_{\rm eff}$.

The possible developments on this work includes the relaxation of the main limitation of the results, for instance the assumption that the noise is independent of the model parameters. Also, we plan to show that the approach presented in the first part of this work can lead to quantitative evaluations in non-Gaussian cases as well, when observables other than the first two moments are also considered.

ACKNOWLEDGMENTS

We thank Baptiste Schubnel for interesting discussions and a careful reading of the manuscript. We also wish to thank the anonymous referee for his detailed and useful comments on the manuscript. JC acknowledges the support of the Swiss National Science Foundation. AA is supported by the Zwicky Fellowship at ETH Zurich.

REFERENCES

Albrecht A. et al., 2006, ArXiv (astro-ph/0609591)
Amara A., Réfrégier A., 2007, MNRAS, 381, 1018
Bacon D. J., Goldberg D. M., Rowe B. T. P., Taylor A. N., 2006, MNRAS, 365, 414
Bartelmann M., Schneider P., 2001, Phys. Rep., 340, 291
Bernstein G. M., 2009, ApJ, 695, 652
Bhatia R., 2007, Positive Definite Matrices. Princeton Univ. Press, Princeton, NJ
Bridle S. L., Hobson M. P., Lasenby A. N., Saunders R., 1998, MNRAS, 299, 895
Caticha A., 2008, ArXiv (0808.0012)
Chevallier M., Polarski D., 2001, Int. J. Mod. Phys., D10, 213
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Er X., Li G., Schneider P., 2010, ArXiv (astro-ph/1008.3088)
Fisher R. A., 1925, Proc. Cambridge Philos. Soc., 22, 700
Hu W., Jain B., 2004, Phys. Rev. D, 70, 043009
Hu W., Tegmark M., 1999, ApJ, 514, L65
Jaynes E., 1983, Papers on Probability, Statistics and Statistical Physics. Reidel, Dordrecht
Jaynes E. T., 1957, Phys. Rev., 106, 620
Jaynes E. T., Bretthorst G. L., 2003, Probability Theory: The Logic of Science. Cambridge Univ. Press, Cambridge
Jennrich R. I., Moore R. H., 1975, Proc. Statistical Comput., American Statistical Association, Washington, p. 57
Kaiser N., Squires G., 1993, ApJ, 404, 441
Kopman B. O., 1936, Trans. American Math. Soc., 39, 399

Kullback S., 1959, Information Theory and Statistics. Wiley, New York
Linder E. V., 2003, Phys. Rev. Lett., 90, 091301
Maisinger K., Hobson M. P., Lasenby A. N., 2004, MNRAS, 347, 339
Marshall P. J., Hobson M. P., Gull S. F., Bridle S. L., 2002, MNRAS, 335, 1037
Massey R., Rowe B., Refregier A., Bacon D. J., Bergé J., 2007, MNRAS, 380, 229
Munshi D., Valageas P., Van Waerbeke L., Heavens A., 2006, Phys. Rep., 462, 67
Okura Y., Umetsu K., Futamase T., 2007, ApJ, 660, 995
Parkinson D., Blake C., Kunz M., Bassett B. A., Nichol R. C., Glazebrook K., 2007, MNRAS, 377, 185
Pires S., Amara A., 2010, ApJ, 723, 1507
Pitman E. J. G., Wishart J., 1936, in Proc. Cambridge Philos. Soc. Vol. 32, Sufficient Statistics and Intrinsic Accuracy. Cambridge Philos. Soc., Cambridge, p. 567
Rao C., 1973, Lineare statistische Methoden und ihre Anwendungen. Akademie Verlag Berlin
Refregier A., 2003, ARA&A, 41, 645
Refregier A., Amara A., Kitching T., Rassat A., 2011, A&A, 528, 33
Schneider P., Ehlers J., Falco E. E., 1992, Gravitational Lenses. Springer-Verlag, Berlin
Schneider P., Er X., 2008, A&A, 485, 363
Schneider P., Kochanek C. S., Wambsganss J., 2006, Gravitational Lensing: Strong, Weak and Micro. Springer, Heidelberg
Seitz S., Schneider P., 2001, A&A, 374, 740
Shannon C. E., 1948, Bell Syst. Technical J., 27, 379
Skilling J., Bryan R. K., 1984, MNRAS, 211, 111
Smith R. E. et al., 2003, MNRAS, 341, 1311
Starck J., Pantin E., 1996, Vistas Astron., 40, 563
Taylor A. N., Watts P. I. R., 2001, MNRAS, 328, 1027
Tegmark M., 1997, Phys. Rev. Lett., 79, 3806
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
The Dark Energy Survey Collaboration 2005, ArXiv (astro-ph/0510346)
Vallinotto A., Dodelson S., Zhang P., 2010, ArXiv (astro-ph/1009.5590)
van den Bos A., 2007, Parameter Estimation for Scientists and Engineers. Wiley, New York
Velander M., Kuijken K., Schrabback T., 2011, MNRAS, 412, 26665
Zografos K., Ferentinos K., 1994, Metrika, 41, 109
Zunckel C., Trotta R., 2007, MNRAS, 380, 865

APPENDIX A: CRAMÉR–RAO INEQUALITY

In this Appendix, we provide a unified derivation of the Cramér–Rao inequality in the multidimensional case [based on Rao (1973)] and its relation to maximum entropy distributions. We denote the vector of model parameters of dimension n by

$$\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_n\right) \tag{A1}$$

and a vector of functions of dimension m, the estimators

$$\hat{\boldsymbol{D}} = (\hat{D}_1, \ldots, \hat{D}_m), \tag{A2}$$

with expectation values $D_i(\boldsymbol{\alpha}) = \langle \hat{D}_i(x) \rangle$. In the following, we rely on Gram matrices, whose elements are defined by scalar products. Namely, for a set of vectors $\boldsymbol{y}_i$, the Gram matrix $\mathbf{Y}$ generated by this set of vectors is defined as

$$\mathbf{Y}_{ij} = \boldsymbol{y}_i \cdot \boldsymbol{y}_j. \tag{A3}$$

Gram matrices are positive definite and have the same rank as the set of vectors that generate them. Especially, if the vectors are linearly independent, the Gram matrix is strictly positive definite and invertible.

We adopt a vectorial notation for functions, writing scalar products between vectors as

$$f \times g \equiv \int \mathrm{d}x \; p_X(x, \boldsymbol{\alpha}) f(x) g(x), \tag{A4}$$

with $p_X(x, \boldsymbol{\alpha})$ being the probability density function of the variable X of interest. In this notation, both the Fisher information matrix and the covariance matrix are seen to be Gram matrices. We have that the Fisher information matrix reads

$$F_{\alpha_i \alpha_j} = f_{\alpha_i} \times f_{\alpha_j}, \quad f_{\alpha_i}(x) = \frac{\partial \ln p_X(x, \boldsymbol{\alpha})}{\partial \alpha_i}, \tag{A5}$$

while the covariance matrix of the estimators is

$$C_{ij} = g_i \times g_j, \quad g_i(x, \boldsymbol{\alpha}) = \hat{D}_i(x) - D_i(\boldsymbol{\alpha}). \tag{A6}$$

For simplicity and since it is sufficiently generic for our purpose, we will assume that both sets of vectors $\boldsymbol{f}$ and $\boldsymbol{g}$ are linearly independent, so that both matrices can be inverted. Note that we also have

$$\frac{\partial D_i}{\partial \alpha_j} = \int \mathrm{d}x \; p_X(x, \boldsymbol{\alpha}) \, \hat{D}_i(x) \frac{\partial \ln p_X(x, \boldsymbol{\alpha})}{\partial \alpha_j} = g_i \times f_{\alpha_j}. \tag{A7}$$

The Gram matrix $\mathbf{G}$ of dimension $[(m+n) \times (m+n)]$ generated by the set of vectors $(g_1, \ldots, g_m, f_{\alpha_1}, \ldots, f_{\alpha_n})$ takes the form

$$\mathbf{G} = \begin{pmatrix} \mathbf{C} & \boldsymbol{\Delta} \\ \boldsymbol{\Delta}^{\mathrm{T}} & \mathbf{F} \end{pmatrix}, \quad \Delta_{i\alpha_j} = g_i \times f_{\alpha_j}, \tag{A8}$$

and is also positive definite due to its very definition. It is congruent to the matrix

$$\mathbf{Y}\mathbf{G}\mathbf{Y}^{\mathrm{T}} = \begin{pmatrix} \mathbf{C} - \boldsymbol{\Delta}\mathbf{F}^{-1}\boldsymbol{\Delta}^{\mathrm{T}} & 0 \\ 0 & \mathbf{F} \end{pmatrix} \tag{A9}$$

with

$$\mathbf{Y} = \begin{pmatrix} 1_{m\times m} & -\boldsymbol{\Delta}\mathbf{F}^{-1} \\ 0 & 1_{n\times n} \end{pmatrix}. \tag{A10}$$

Since two congruent matrices have the same number of positive, zero and negative eigenvalues, respectively, and since both $\mathbf{F}$ and $\mathbf{G}$ are positive, we can conclude that

$$\mathbf{C} \geq \boldsymbol{\Delta}\mathbf{F}^{-1}\boldsymbol{\Delta}^{\mathrm{T}}, \tag{A11}$$

which is the Cramér–Rao inequality. The lower bound on the amount of information is seen from the fact that for any matrix written in block form holds

$$\begin{pmatrix} \mathbf{C} & \boldsymbol{\Delta} \\ \boldsymbol{\Delta}^{\mathrm{T}} & \mathbf{F} \end{pmatrix} \geq 0 \Leftrightarrow \begin{pmatrix} \mathbf{F} & \boldsymbol{\Delta}^{\mathrm{T}} \\ \boldsymbol{\Delta} & \mathbf{C} \end{pmatrix} \geq 0 \tag{A12}$$

and using the same congruence argument leads to the lower bound on information:

$$\mathbf{F} \geq \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Delta}. \tag{A13}$$

Assume now that we have a probability density function such that this inequality is in fact an equality, i.e.

$$\mathbf{F} = \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Delta}. \tag{A14}$$

By the above argument, the Gram matrix generated by

$$(f_{\alpha_1}, \ldots, f_{\alpha_n}, g_1, \ldots, g_m) \tag{A15}$$

is congruent to the matrix

$$\begin{pmatrix} 0_{n\times n} & 0 \\ 0 & \mathbf{C} \end{pmatrix} \tag{A16}$$

and has rank m. By assumption, the covariance matrix is invertible, such that the set $(g_1, \ldots, g_m)$ alone has rank m. It implies that each

of the $\boldsymbol{f}$ vectors can be written as a linear combination of the g vectors,

$$\boldsymbol{f}_{\alpha_i} = \sum_{j=1}^{m} A_j \boldsymbol{g}_j, \tag{A17}$$

or, more explicitly,

$$\frac{\partial \ln p_X(x, \boldsymbol{\alpha})}{\partial \alpha_i} = \sum_{j=1}^{m} A_j(\boldsymbol{\alpha})[\hat{D}_j(x) - D_j(\boldsymbol{\alpha})], \tag{A18}$$

where the key point is that the coefficients A_j are independent of x. Integrating this equation, we obtain

$$\ln p_X(x, \boldsymbol{\alpha}) = -\sum_{i=1}^{m} \lambda_i(\boldsymbol{\alpha})\hat{D}_i(x) - \ln Z(\boldsymbol{\alpha}) + \ln q_X(x) \tag{A19}$$

for some functions λ and Z of the model parameters only, and a function q_X of x only. We obtain thus

$$p_X(x, \boldsymbol{\alpha}) = \frac{q_X(x)}{Z(\boldsymbol{\alpha})} \exp\left(-\sum_{i=1}^{m} \lambda_i(\boldsymbol{\alpha})\hat{D}_i(x) \right). \tag{A20}$$

This is precisely the distribution that we obtain by maximizing the entropy relative to $q_X(x)$, while satisfying the constraints

$$D_i(\boldsymbol{\alpha}) = \langle \hat{D}_i(x) \rangle, \quad i = 1, \ldots, m. \tag{A21}$$

Taking q_X as the uniform distribution makes it identical to the formula in equation (24).

APPENDIX B: POINT FIELDS

The data consist of a set of numbers at each position where a galaxy sits and a measurement was done. We use the handy notation in terms of Dirac delta function,

$$\phi(\boldsymbol{x}) = \sum_i \epsilon_i \delta^{\mathrm{D}}(\boldsymbol{x} - \boldsymbol{x}_i), \tag{B1}$$

where the sum runs over the positions $\boldsymbol{x}_i$ for which ϵ is measured. To obtain the spectral matrices, we need the Fourier transform of the field, which reads in our case

$$\tilde{\phi}(\boldsymbol{l}) = \sum_i \epsilon_i \exp(-i\boldsymbol{l} \cdot \boldsymbol{x}_i). \tag{B2}$$

In this work, we assume that the set of points shows negligible clustering so that the probability density function for the joint occurrence of a particular set of galaxy positions is uniform.

We decompose in the following the wavevector $\boldsymbol{k}$ on the flat sky in terms of its modulus and polar angle as

$$\boldsymbol{l} = l \begin{pmatrix} \cos\phi_l \\ \sin\phi_l \end{pmatrix}. \tag{B3}$$

B1 Ellipticities

When the two ellipticity components are measured, we have two such fields ϕ_1, ϕ_2 at our disposal. For instance, the field describing the first component becomes

$$\tilde{\phi}_1(\boldsymbol{l}) = \sum_i \epsilon_i^1 \exp(-i\boldsymbol{l} \cdot \boldsymbol{x}_i). \tag{B4}$$

We assume that the measured ellipticities trace the shear fields in the sense that the measured components are built out of the shear at that position plus some value unrelated to it:

$$\begin{aligned} \epsilon_i^1 &= \gamma_1(\boldsymbol{x}_i) + \epsilon^1_{\mathrm{int},\,i}, \\ \epsilon_i^2 &= \gamma_2(\boldsymbol{x}_i) + \epsilon^2_{\mathrm{int},\,i}. \end{aligned} \tag{B5}$$

The vector $\boldsymbol{v}$ relating the spectral matrices of the ellipticities and the convergence is then obtained by combining (B4) with the above relations (B5) in its definition (69), and using the relation between shears and convergence in equation (82). The result is

$$\boldsymbol{v} = \bar{n}_\gamma \begin{pmatrix} \cos 2\phi_l \\ \sin 2\phi_l \end{pmatrix}, \tag{B6}$$

where $\bar{n}_\gamma$ is the number density of galaxies for which ellipticity measurements are available. Under our assumptions of uncorrelated intrinsic ellipticities, with dispersions of equal magnitude σ_γ^2 for the two components, the noise matrix $\mathbf{N}$ becomes

$$\mathbf{N} = \bar{n}_\gamma \begin{pmatrix} \sigma_\gamma^2 & 0 \\ 0 & \sigma_\gamma^2 \end{pmatrix}. \tag{B7}$$

The effective noise, given in equation (71), is readily computed:

$$N_{\mathrm{eff}}^\gamma = \frac{\sigma_\gamma^2}{\bar{n}_\gamma}. \tag{B8}$$

B2 Sizes

As has been noted in the main text, the apparent sizes of galaxies are modified by lensing as

$$s^i_{\mathrm{obs}} = s^i_{\mathrm{int}}(1 + \alpha_{\mathrm{s}}\kappa) \tag{B9}$$

for some coefficient α_{s} which is unity in pure weak lensing theory. Denoting the number of galaxies for which sizes measurements are available by n_{s}, and the mean intrinsic size of the sample by $\bar{s}_{\mathrm{int}}$, the spectrum of the size field reduces, under the assumption of uncorrelated intrinsic sizes, to

$$C^{\mathrm{s}}(l) = \bar{n}_{\mathrm{s}}^2 \bar{s}_{\mathrm{int}}^2 \alpha_{\mathrm{s}}^2 C^\kappa(l) + \bar{n}_{\mathrm{s}} \sigma_{\mathrm{s}}^2. \tag{B10}$$

The vector $\boldsymbol{v}$ and matrix $\mathbf{N}$ are now numbers that are read out from the above equation as

$$\begin{aligned} \boldsymbol{v} &= \bar{n}_{\mathrm{s}} \bar{s}_{\mathrm{int}} \alpha_{\mathrm{s}}, \\ \mathbf{N} &= \bar{n}_{\mathrm{s}} \sigma_{\mathrm{s}}^2, \end{aligned} \tag{B11}$$

leading to the effective noise

$$N_{\mathrm{eff}}^{\mathrm{s}} = \frac{1}{\alpha_{\mathrm{s}}^2} \left(\frac{\sigma_{\mathrm{s}}}{\bar{s}_{\mathrm{int}}} \right)^2 \frac{1}{\bar{n}_{\mathrm{s}}}. \tag{B12}$$

B3 Second order, flexion

Denoting by $\bar{n}_{\mathcal{F}}$ and $\bar{n}_{\mathcal{G}}$ the number of galaxies for which $\mathcal{F}$ and $\mathcal{G}$ are measured, the vectors linking the flexion to convergence are

$$\boldsymbol{v}^{\mathcal{F}} = -il\bar{n}_{\mathcal{F}} \begin{pmatrix} \cos\phi_l \\ \sin\phi_l \end{pmatrix} \tag{B13}$$

and

$$\boldsymbol{v}^{\mathcal{G}} = -il\bar{n}_{\mathcal{G}} \begin{pmatrix} \cos 3\phi_l \\ \sin 3\phi_l \end{pmatrix}. \tag{B14}$$

Using again the assumption of uncorrelated intrinsic components, we have the four-dimensional diagonal noise matrix

$$\mathbf{N} = \begin{pmatrix} \bar{n}_{\mathcal{F}} \sigma_{\mathcal{F}}^2 \times 1_{2x2} & 0 \\ 0 & \bar{n}_{\mathcal{G}} \sigma_{\mathcal{G}}^2 \times 1_{2x2} \end{pmatrix}, \tag{B15}$$

leading to the effective noise, this time mode-dependent:

$$\frac{1}{N_{\mathrm{eff}}^{\mathcal{FG}}} = l^2 \left(\frac{\bar{n}_F}{\sigma_F^2} + \frac{\bar{n}_G}{\sigma_G^2} \right). \tag{B16}$$

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