Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 14 September 2014 (MN LATEX style file v2.2)

Generalised Fisher Matrices

A.F. Heavens[1⋆], M. Seikel[2], B.D. Nord[3], M. Aich[4], Y. Bouffanais[1], B.A. Bassett[5,6,7], M.P. Hobson[8]

[1] *Imperial Centre for Inference and Cosmology, Department of Physics, Imperial College, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, U.K.*
[2] *The UCT Astrophysics, Cosmology and Gravity Centre, Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch 7701, Cape Town, South Africa*
[3] *Department of Physics, University of Michigan, Ann Arbor, Michigan, United States*
[4] *School of Mathematics, Statistics & Computer Science, University of KwaZulu-Natal, Durban 4000, South Africa*
[5] *African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg, 7945, South Africa*
[6] *Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch, Cape Town, 7700, South Africa*
[7] *South African Astronomical Observatory, Observatory Road, Observatory, Cape Town, 7935, South Africa*
[8] *Battcock Centre for Experimental Astrophysics, University of Cambridge, Madingley Road, Cambridge, CB3 0HA*

Accepted ; Received ; in original form

ABSTRACT

The Fisher Information Matrix formalism (Fisher 1935) is extended to cases where the data is divided into two parts $(\mathbf{X}, \mathbf{Y})$, where the expectation value of $\mathbf{Y}$ depends on $\mathbf{X}$ according to some theoretical model, and $\mathbf{X}$ and $\mathbf{Y}$ both have errors with arbitrary covariance. In the simplest case, $(\mathbf{X}, \mathbf{Y})$ represent data pairs of abscissa and ordinate, in which case the analysis deals with the case of data pairs with errors in both coordinates, but $\mathbf{X}$ can be *any* measured quantities on which $\mathbf{Y}$ depends. The analysis applies for arbitrary covariance, provided all errors are gaussian, and provided the errors in $\mathbf{X}$ are small, both in comparison with the scale over which the expected signal $\mathbf{Y}$ changes, and with the width of the prior distribution. This generalises the Fisher Matrix approach, which normally only considers errors in the 'ordinate' $\mathbf{Y}$. In this work, we include errors in $\mathbf{X}$ by marginalising over latent variables, effectively employing a Bayesian hierarchical model, and deriving the Fisher Matrix for this more general case. The methods here also extend to likelihood surfaces which are not gaussian in the parameter space, and so techniques such as DALI (Derivative Approximation for Likelihoods) can be generalised straightforwardly to include arbitrary gaussian data error covariances. For simple mock data and theoretical models, we compare to Markov Chain Monte Carlo experiments, illustrating the method with cosmological supernova data. We also include the new method in the Fisher4Cast software.

Key words: statistics: general — statistics: Fisher matrix — cosmology: forecasts

1 INTRODUCTION

The Fisher Information Matrix or simply Fisher Matrix has become one of the most widely used statistical tools for forecasting the errors in parameter estimation problems. It provides lower limits on the variances (through the Cramér-Rao inequality), and the expected covariances of estimates of model parameters from maximum likelihood, or maximum posterior, techniques, for a given experimental design. If we further assume gaussianity in two respects: that the data are jointly gaussian-distributed, and that the posterior for the parameters is gaussian, then the Fisher matrix determines the full expected posterior. For data pairs $\{X_i, Y_i\}$ with no errors in X, the problem was solved many years ago (Fisher 1935). The main value of the Fisher matrix technique is in being able to obtain error forecasts without any data, real or simulated, and is generally much faster than computing full posterior distributions with simulations (Acquaviva et al. 2012; Bassett et al. 2009). It is however only a first step, as it assumes the posteriors are well described by multivariate gaussian distributions, and this may not hold (e.g., Wolz

⋆ e-mail: a.heavens@imperial.ac.uk

et al. 2012), when more sophisticated analysis may be required, but it is still a very valuable tool for experimental design. Furthermore, more sophisticated forecasts for likelihood surfaces which are non-gaussian in the parameter space now exist (Sellentin et al. 2014).

From the initial derivations of the Fisher Matrix in the cosmological context (Vogeley & Szalay 1996; Tegmark, Taylor, & Heavens 1997), we have arrived today at very mature applications and implementations (e.g., Bassett et al. 2009; Coe 2009; Refregier et al. 2011). The Fisher Matrix has been useful in proposals and projections for surveys, such as for the Cosmic Microwave Background (Taylor et al. 1997), spectroscopic galaxy surveys (Schlegel et al. 2011), the Dark Energy Survey (DES Collaboration 2005), large-scale structure (Cunha 2009), and in the broader discussion of the investigation of Dark Energy (Albrecht et al. 2006) and estimation of neutrino masses with the future European Space Agency Euclid mission (Kitching et al. 2008).

For the purposes of review and later reference in this work, we summarise the basic Fisher Matrix formalism. We begin with the likelihood of a set of data, $\mathbf{d}$ given (or conditional upon) a set of model parameters, represented by a vector $\boldsymbol{\theta}$: $p(\mathbf{d}|\boldsymbol{\theta})$. In the simplest case, $\mathbf{d}$ represents only the ordinates, $\mathbf{Y}$. Later in the paper, we will take it to be the union of the ordinates and any other measured quantities on which $\mathbf{Y}$ depends, such as abscissa values, and which may be subject to error. In practice what is typically required is the posterior distribution of $\boldsymbol{\theta}$, given the data $\mathbf{d}$. Assuming an uninformative prior on the parameters, $p(\boldsymbol{\theta}) =$ constant, Bayes' Theorem implies $p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta}) = L$, the likelihood. The log-likelihood is then Taylor-expanded about its maximum. The first term is a constant, irrelevant for the discussion of parameter constraint forecasts; the second term is the first derivative, which vanishes at the point of maximum likelihood; the third term is the Hessian (curvature matrix) of the likelihood, and is the term whose ensemble average (over the data) gives the Fisher Matrix:

$$F_{\alpha\beta} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle , \tag{1}$$

where α and β label the parameters. For the case of a gaussian likelihood, this is analytically computable, and can depend only on the expectation values of the data, $\mu(\boldsymbol{\theta}) \equiv \langle \mathbf{d}(\boldsymbol{\theta})\rangle$, and the covariance, $\mathsf{C}(\boldsymbol{\theta}) \equiv \langle (\mathbf{d}-\boldsymbol{\mu})^T(\mathbf{d}-\boldsymbol{\mu})\rangle$. This results in the following form for the Fisher Matrix (Tegmark, Taylor, & Heavens 1997).

$$\mathsf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[\mathsf{C}^{-1}\mathsf{C}_{,\alpha}\mathsf{C}^{-1}\mathsf{C}_{,\beta} + \mathsf{C}^{-1}(\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}_{,\beta}^T + \boldsymbol{\mu}_{,\beta}\boldsymbol{\mu}_{,\alpha}^T)\right] . \tag{2}$$

An early example of dealing with errors in both variables was straight-line fitting, where both the statistics and astronomy communities used either *ad hoc* choices for the axis, or ultimately arbitrary combinations e.g., the bisector or the average of the one-dimensional fits on either axis. The evolution to two-dimensional or joint-distribution fitting was accompanied by a slow transition to the Bayesian perspective (Gull 1989). New tools for fitting data in the presence of two-dimensional errors have been developed and used to extract improved cosmological constraints from supernovae populations (March et al. 2011). Here, we develop the application of two-dimensional errors in the predictive Fisher Matrix formalism itself, but the formalism can treat more general cases where the signal depends on arbitrary extra parameters . For pedagogical discussions of straight-line fitting and Bayesian approaches to fitting, see for example Hogg et al. (2010); D'Agostini (2005); Kelly (2011).

The remainder of the paper is organized as follows: §2 describes the formal derivation of the generalized Fisher matrix for the case of dependence of $\mathbf{Y}$ on an arbitrary set of gaussian-distributed variables $\mathbf{X}$; §3 describes an application of this formalism to a particular experiment, with tests on simulated data. We present conclusions in §4. For the reader who is interested only in the application of the result, this is effected by simply replacing the covariance matrix C in equation (2) by the matrix R computed in equation (18).

2 FORMALISM OF THE EXTENSION

Throughout this paper, we follow the formalism and notation of Bassett et al. (2009). In this method, we use a Taylor expansion of the log-likelihood, and derive the generalised Fisher Matrix from first principles. The general aim is to find an expression for the Fisher Matrix for an experiment with gaussian errors in $\mathbf{X}$ and $\mathbf{Y}$, arbitrary correlations of errors (i.e. errors in Y_i can be correlated with errors in X_j, for any i, j). As previously mentioned, the formalism covers the case when $\mathbf{X}$ represents the abscissa values of the data points, but it need not, and the extra variables may not be associated with an individual Y_i at all.

2.1 General Method with X-Y Covariance

Let the set of measurements be $\{X_i\}, \{Y_j\}$, with $i = 1, \ldots M$ and $j = 1, \ldots N$. In the simplest case, $M = N$ and the dataset is a set of (X, Y) data pairs, but this is not necessary; all that is required is that there a model which returns the expectation value of $\mathbf{Y}$ as a function of $\mathbf{X}$, and which in general will depend also on some model parameters, represented collectively by $\boldsymbol{\theta}$, being a vector θ_α with $\alpha = 1, \ldots P$. It is the posterior probability of $\boldsymbol{\theta}$ which we wish to calculate. We give an example later.

We assume $\mathbf{X}$ and $\mathbf{Y}$ have Gaussian errors, around true values $\mathbf{x}$, $\mathbf{y}$, with a covariance matrix C. $\mathbf{x}$ and $\mathbf{y}$ are not directly

observed. This amounts to a hierarchical model, where the observables $\mathbf{X}, \mathbf{Y}$ depend on some unobservable latent variables $\mathbf{x}$, which are essentially nuisance parameters. The $\mathbf{y}$ are not independent nuisance parameters as they are assumed to be related precisely by a theoretical model $\mathbf{y} = \boldsymbol{\mu}(\mathbf{x})$, which also depends on $\boldsymbol{\theta}$. We seek the posterior $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})$. With a uniform prior for $\boldsymbol{\theta}$, this is proportional to the likelihood $L = p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})$. We write this as the marginalised distribution over $\mathbf{x}$ and $\mathbf{y}$ as

$$\begin{aligned} L &= \int p(\mathbf{X}, \mathbf{Y}, \mathbf{x}, \mathbf{y}|\boldsymbol{\theta})\, d\mathbf{x}\, d\mathbf{y} = \int p(\mathbf{X}, \mathbf{Y}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})\, d\mathbf{x}\, d\mathbf{y} \\ &= \int p(\mathbf{X}, \mathbf{Y}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}\, d\mathbf{y} \end{aligned} \tag{3}$$

where we have expanded the condition to include the latent variables, and then further expanded the condition of $p(\mathbf{y})$ to include $\mathbf{x}$.

We integrate over $\mathbf{y}$ using a delta function, $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \delta(\mathbf{y} - \boldsymbol{\mu}(\mathbf{x}))$, and assume for now a uniform prior for $\mathbf{x}$:

$$L = \int p(\mathbf{X}, \mathbf{Y}|\mathbf{x}, \boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\theta})\, d^M\mathbf{x}. \tag{4}$$

At the cost of some algebraic complexity, we can introduce an informative prior (parent distribution) for $\mathbf{x}$. In Appendix B, we generalise the analysis by assuming a gaussian population prior $p(\mathbf{x})$, and show that we recover the simpler result obtained in the main text in the limit that the errors in $\mathbf{x}$ are small enough that the prior can be considered constant across the error range of individual data points. Note that formally we assume the prior is independent of the model parameters, but in the limit discussed in the main text, any such dependence does not affect the result. See Gull (1989) and Kelly (2011) for further discussion of these points. In this paper we are not explicitly concerned with biases, but it is important to note Gull's point that estimates of parameters, such as the slope of a straight-line fit with errors in both coordinates, will be biased, even with an informative prior, unless the width of the prior is a hyperparameter that is marginalized over. No doubt similar considerations will be important in applications of the more complicated situation considered here.

Next, we make the critical assumption that we can truncate at the linear term of the Taylor expansion of $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}(\mathbf{x}) = \boldsymbol{\mu}(\mathbf{X}) + \mathsf{T}(\mathbf{X})\,(\mathbf{x} - \mathbf{X}), \tag{5}$$

where

$$\mathsf{T}_{ij} \equiv \left.\frac{\partial \mu_i}{\partial x_j}\right|_{\mathbf{x}=\mathbf{X}}. \tag{6}$$

In the case when $\mathbf{X}$ represents the abscissa values, we would expect T to be diagonal.

We are essentially assuming that the function $\boldsymbol{\mu}(\mathbf{x})$ is linear across the width of the gaussian error distribution of $\mathbf{x}$, and this allows the likelihood to be integrated analytically, as it is simply a gaussian integral:

$$L \propto \int \frac{1}{\sqrt{\det \mathsf{C}}} \exp\left(-\frac{Q}{2}\right) d\mathbf{x} \tag{7}$$

where $Q \equiv (\mathbf{Z} - \mathbf{z})^T \mathsf{C}^{-1} (\mathbf{Z} - \mathbf{z})$, and $\mathbf{z}$ and $\mathbf{Z}$ are $M + N$-dimensional vectors: $z_i = x_i$ and $Z_i = X_i$ for $i \leqslant M$, $Z_{M+j} = Y_j$ and $z_{M+j} = \mu_j(\mathbf{X}) + [\mathsf{T}(\mathbf{X})(\mathbf{x} - \mathbf{X})]_j$.

The covariance matrix of the data can be written in block form as

$$\mathsf{C} = \begin{array}{c} \\ X \\ Y \end{array} \begin{array}{c} \begin{array}{cc} X & Y \end{array} \\ \begin{pmatrix} \mathsf{C}_{XX} & \mathsf{C}_{XY} \\ \mathsf{C}_{XY}^T & \mathsf{C}_{YY} \end{pmatrix} \end{array}. \tag{8}$$

Note that C_{XY} is not symmetrical, nor invertible or even square in general; although C_{XX} and C_{YY} are. The covariance matrix may include a number of elements, such as intrinsic scatter and measurement noise, with individual covariance matrices adding to give the final C. The inverse of C is

$$\mathsf{C}^{-1} = \begin{pmatrix} \mathsf{G} & -\mathsf{H} \\ -\mathsf{H}^T & \mathsf{E} \end{pmatrix} \tag{9}$$

where

$$\begin{aligned} \mathsf{G} &= \mathsf{C}_{XX}^{-1} + \mathsf{C}_{XX}^{-1} \mathsf{C}_{XY} \mathsf{E} \mathsf{C}_{XY}^T \mathsf{C}_{XX}^{-1} \\ \mathsf{H} &= \mathsf{C}_{XX}^{-1} \mathsf{C}_{XY} \mathsf{E} \\ \mathsf{E} &= (\mathsf{C}_{YY} - \mathsf{C}_{XY}^T \mathsf{C}_{XX}^{-1} \mathsf{C}_{XY})^{-1}. \end{aligned} \tag{10}$$

Defining $\tilde{\mathbf{x}} \equiv \mathbf{X} - \mathbf{x}$, and $\tilde{\mathbf{Y}} \equiv \mathbf{Y} - \boldsymbol{\mu}(\mathbf{X})$, we collect together the terms as follows:

$$Q = \tilde{\mathbf{x}}^T \mathsf{G} \tilde{\mathbf{x}} + (\tilde{\mathbf{Y}} + \mathsf{T}\tilde{\mathbf{x}})^T \mathsf{E} (\tilde{\mathbf{Y}} + \mathsf{T}\tilde{\mathbf{x}}) - \tilde{\mathbf{x}}^T \mathsf{H} (\tilde{\mathbf{Y}} + \mathsf{T}\tilde{\mathbf{x}}) - (\tilde{\mathbf{Y}} + \mathsf{T}\tilde{\mathbf{x}})^T \mathsf{H}^T \tilde{\mathbf{x}}. \tag{11}$$

Q has the quadratic form

$$Q = \tilde{\mathbf{x}}^T \mathsf{A} \tilde{\mathbf{x}} - \mathbf{B}^T \tilde{\mathbf{x}} - \tilde{\mathbf{x}}^T \mathbf{B} + Q', \tag{12}$$

where

$$\begin{aligned} \mathsf{A} &= \mathsf{G} + \mathsf{T}^T\mathsf{E}\mathsf{T} - \mathsf{H}\mathsf{T} - \mathsf{T}^T\mathsf{H}^T \\ \mathbf{B} &= (\mathsf{H} - \mathsf{T}^T\mathsf{E})\tilde{\mathbf{Y}} \equiv \mathsf{P}\tilde{\mathbf{Y}} \\ Q' &= \tilde{\mathbf{Y}}^T\mathsf{E}\tilde{\mathbf{Y}}. \end{aligned} \tag{13}$$

With the definition of Q in Eqn. 12, the gaussian integral of Eqn. 7 can be performed, using

$$\int \exp\left(-\frac{1}{2}\tilde{\mathbf{x}}^T\mathsf{A}\tilde{\mathbf{x}} + \mathbf{B}^T\tilde{\mathbf{x}}\right) d\tilde{\mathbf{x}} = \frac{(2\pi)^{N/2}}{\sqrt{\det A}} \exp\left(\frac{1}{2}\mathbf{B}^T\mathsf{A}^{-1}\mathbf{B}\right), \tag{14}$$

and noting that Q' is independent of $\tilde{x}$. The likelihood then simplifies after a few lines of algebra to

$$L \propto \frac{1}{\sqrt{\det\mathsf{A}\det\mathsf{C}}} \exp\left(-\frac{1}{2}\tilde{\mathbf{Y}}^T\mathsf{R}^{-1}\tilde{\mathbf{Y}}\right), \tag{15}$$

where the inverse of the marginal covariance matrix of $\tilde{\mathbf{Y}}$ is

$$\mathsf{R}^{-1} = \mathsf{E} - \mathsf{P}^T\mathsf{A}^{-1}\mathsf{P}. \tag{16}$$

We use the Woodbury formula (Woodbury 1950)

$$(\mathsf{K} + \mathsf{U}\mathsf{W}\mathsf{V})^{-1} = \mathsf{K}^{-1} - \mathsf{K}^{-1}\mathsf{U}(\mathsf{W}^{-1} + \mathsf{V}\mathsf{K}^{-1}\mathsf{U})^{-1}V\mathsf{K}^{-1} \tag{17}$$

to obtain after some algebra

$$\mathsf{R} = \mathsf{C}_{\rm YY} - \mathsf{C}_{\rm XY}^T\mathsf{T}^T - \mathsf{T}\mathsf{C}_{\rm XY} + \mathsf{T}\mathsf{C}_{\rm XX}\mathsf{T}^T, \tag{18}$$

which is the key result of the calculation. We can also simplify the pre-factor, $\det\mathsf{A}\det\mathsf{C} = \det\mathsf{R}$ (see Appendix A for the proof). Thus

$$L \propto \frac{1}{\sqrt{\det\mathsf{R}}} \exp\left(-\frac{1}{2}\tilde{\mathbf{Y}}^T\mathsf{R}^{-1}\tilde{\mathbf{Y}}\right). \tag{19}$$

We see that this looks just like a normal gaussian (in terms of data) likelihood, but with the covariance matrix C ($\mathsf{C}_{\rm YY}$ in our current notation) replaced by R. Hence to compute the Fisher matrix, we can use the standard formula found in Eqn. 2 and Eqn. 15 of Tegmark, Taylor, & Heavens (1997), and simply replace C by R:

$$\mathsf{F}_{\alpha\beta} = \frac{1}{2}{\rm Tr}\left[\mathsf{R}^{-1}\mathsf{R}_{,\alpha}\mathsf{R}^{-1}\mathsf{R}_{,\beta} + \mathsf{R}^{-1}(\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}_{,\beta}^T + \boldsymbol{\mu}_{,\beta}\boldsymbol{\mu}_{,\alpha}^T)\right]. \tag{20}$$

This is the main result of this paper. Note that R depends not only on the standard covariance, but also on the covariance in the independent variable, $\mathsf{C}_{\rm XX}$, the meta-covariance, $\mathsf{C}_{\rm XY}$, and the first partial derivatives of the model function $\boldsymbol{\mu}$. In the case of uncorrelated data pairs, the result reduces to that found in March et al (2011). For the simple case of no correlations between $\mathbf{X}$ and $\mathbf{Y}$ values $\mathsf{R} = \mathsf{C}_{\rm YY} + \mathsf{T}^T\mathsf{C}_{\rm XX}\mathsf{T}$, and with diagonal covariance matrices $\mathsf{C}_{\rm YY}$ and $\mathsf{C}_{\rm XX}$ we recover the propagation of error result that the variance of $f \equiv Y - \mu(X)$ for each data point is effectively

$$\sigma_f^2 = \sigma_{\rm Y}^2 + \mu'(X)^2\,\sigma_{\rm X}^2, \tag{21}$$

where $\mu' = \partial\mu/\partial x$ and C can be replaced in the standard Fisher expression (2) by a diagonal $N \times N$ matrix with these enhanced entries.

We now briefly make a few key observations. First, when the derivatives of the model function are zero ($\mathsf{T} = 0$), then the latent variable $\mathbf{x}$ has no bearing on R, and we recover the usual formula for the Fisher Matrix: when $\mathsf{T} = 0$, $\mathsf{R} = \mathsf{C}_{\rm YY}$. Also, in the limit of infinitesimal errors in $\mathbf{X}$, we recover the usual Fisher matrix formula. As remarked earlier, if the errors in $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, and in the limit that the errors in $\mathbf{X}$ are small in comparison with the width of the prior, we recover the result obtained from propagation of errors, namely that the variance of $\mathbf{Y}$ is effectively increased from $\sigma_{\rm Y}^2$ to $\sigma_{\rm Y}^2 + \mu'^2\sigma_{\rm X}^2$. Also, although the main focus of the paper has been on the Fisher matrix, the expression for the likelihood itself (equation 19) can be used without the usual interpretation that it is gaussian in the parameter space, to make predictions for the shape of the likelihood surfaces beyond ellipses. Thus the technology of DALI (Sellentin et al. 2014) can be generalized straightforwardly by replacing the data covariance matrix by R. Finally, even if the covariance matrix of the data (the original C, which is $\mathsf{C}_{\rm YY}$) is independent of the parameters, R is not, because in general T does depend on the parameters.

3 EXAMPLE APPLICATION

As an example for illustration, consider the Type 1A supernova Hubble diagram, which consists of data pairs corresponding to the redshift of the host galaxy of each supernova, and its apparent brightness. In the case presented here $\mathbf{X}$ and $\mathbf{Y}$ have the same length, and represent the redshifts and distance moduli of the supernovae. Various corrections, based on colour and the timescale of decline of the light curve ('stretch'), are applied such that these supernovae act as standard candles with a small dispersion of around 10%. Colours and stretch could be added to $\mathbf{X}$, in which case $M \geqslant 3N$, but $\mathbf{X}$ could also include

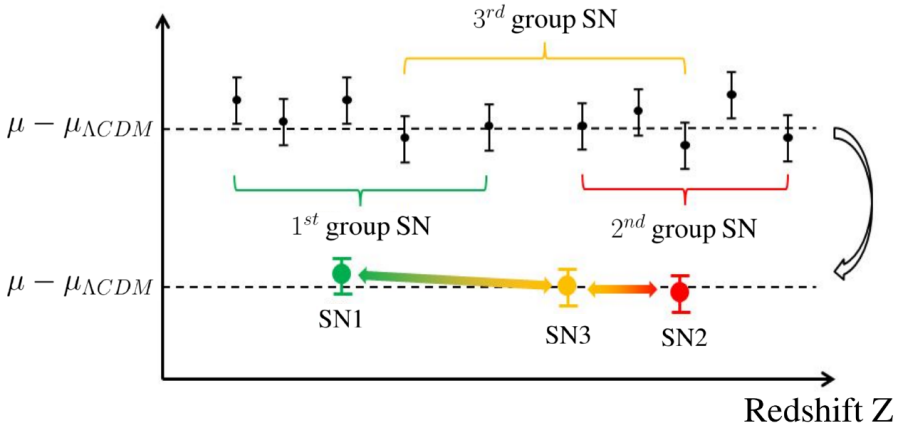

Figure 1. A scenario in which the formation of overlapping weighted combinations of the original data may lead to correlations between X and Y values of different pairs. Here, the Y values have been adjusted to the theoretical curve for a fiducial set of model parameters, which is a function of X, so errors in X propagate into Y, and the weighting then mixes different Y values. This then correlates both X and Y values from different pairs. μ and $\mu_{\Lambda CDM}$ are the measured and theoretical distance moduli, with the theoretical model chosen for illustration to be the ΛCDM concordance model.

variables which are not associated with a single Y_j value (e.g. instrumental calibration). The Λ cold dark matter model plus empirical corrections for colour and stretch then relate $\mathbf{Y}$ to $\mathbf{X}$, dependent on parameters of interest, such as the matter and dark energy content. See Mandel et al. (2011) for a full Bayesian hierarchical model description, and March et al. (2011) for a principled analysis of data, and further discussion of background. Redshift errors obtained from spectroscopy are negligibly small, but if they are photometric redshifts, based on broad-band colours of the host galaxy, then two complications arise. One is that the redshift errors may be large (typically around 5 or 10% for 5-band photometry). The second is that errors in the photometry (such as zero-point errors) will introduce errors in the redshifts, but could also affect the colour corrections for the supernovae themselves. This potentially couples the errors in X and Y for a given data pair. In the rare case of a galaxy with multiple supernovae, a mis-estimation of host galaxy extinction would couple redshift errors, as well as apparent brightness, of the affected supernovae. Kim & Miquel (2007) investigated correlations between redshift and magnitude errors in photometric surveys, and found rather variable correlation coefficients between about 0.35 and 0.95.

A scenario which could couple the errors in X and Y for different data pairs arises if one takes weighted averages of the data. This one might do in order to make the errors closer to gaussian, as we do not know the error distribution for individual supernovae. If this is done with overlapping sub-samples, to maintain a good sampling in redshift (see Fig. 1), then the errors will be coupled. Furthermore, if the Y values are referred to a fiducial model (such as the standard cosmological model), as shown, then this involves dividing by a function of the supernova redshift, which then couples the errors in X to the errors in Y across different (weighted) data pairs. So we see in this example how one can get full covariance between X and Y sets, with non-zero off-diagonal terms of all types.

To illustrate results using the generalised Fisher matrix, we have simulated supernovae with correlated errors in redshift and distance modulus, obtaining an estimate of the posterior for the matter density parameter and cosmological constant, using Markov Chain Monte Carlo techniques. For illustration we show the simplest non-trivial case, where 200 supernovae are drawn from a uniform distribution of redshifts z in the range $0 < z < 1.1$, each having uncorrelated gaussian errors of 0.1 in distance modulus and 0.01 in z; more complicated examples look essentially the same. Fig. 2 shows the comparison of the MCMC error ellipse with the expected error contours from the generalised Fisher Matrix technique, showing good agreement.

4 CONCLUSIONS

In this paper we have considered the Fisher Information Matrix where some subset of the data ($\mathbf{Y}$) depends via a theoretical model $\langle \mathbf{Y} \rangle = \boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\theta})$ on some other set of measured variables ($\mathbf{X}$), and a set of model parameters $\boldsymbol{\theta}$ whose posterior distribution is desired. $\mathbf{X}$ and $\mathbf{Y}$ are assumed to have gaussian errors which can have arbitrary covariance. This includes as a subset the case of (X, Y) data pairs with errors in both coordinates, with correlations between one independent variable and a different dependent variable, but the analysis is more general, and $\mathbf{X}$ can included any other measured quantities. The main result, equation (20), is similar to the standard Fisher matrix, but with the covariance matrix replaced by a more complicated matrix (18) derived from the expanded covariance matrix of all variables, and the partial derivatives of the expected signals with respect to the dependent variables. The result is valid for situations where two conditions hold: the first is that a Taylor

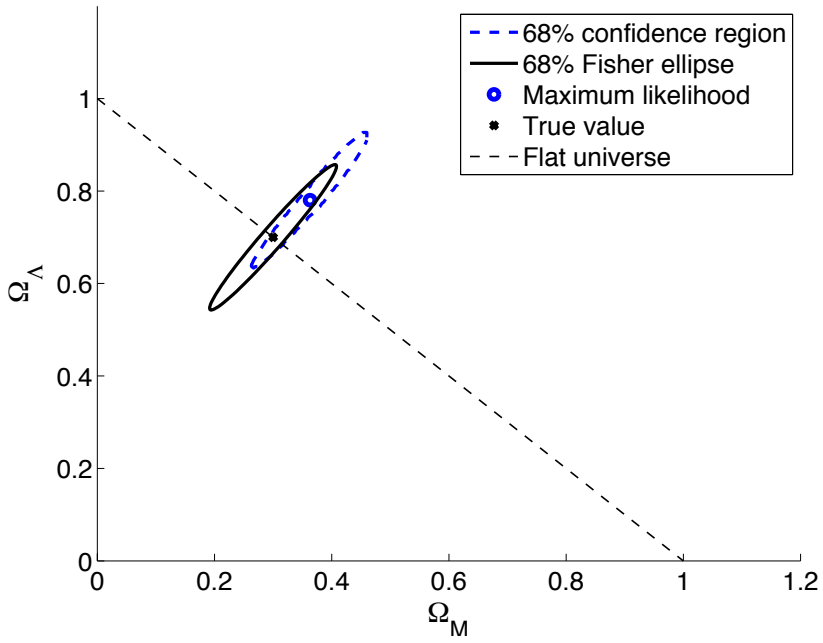

Figure 2. Generalised Fisher Matrix calculations compared with MCMC results from simulated supernova data generated with correlations between **X** and **Y** values in each data pair. The likelihood is accurately a bivariate gaussian for this example, and there is good agreement in the shape, size and orientation of the ellipses, with the actual likelihood offset from the true solution in accordance with expectation.

expansion of the expected signal to linear order is valid across the gaussian error of the independent variables; the second is that the errors in the independent variables are small compared with the width of the prior distribution. At the price of some complexity, we present a perturbative correction when the latter condition does not hold. In the case when the errors are uncorrelated between data pairs, the result reduces to the result one obtains from propagation of errors, where the variance of the dependent variable is increased from $\sigma_{\rm Y}^2$ to $\sigma_{\rm Y}^2 + (\partial\mu/\partial x)^2\sigma_{\rm X}^2$. Since we compute the likelihood itself, it may be used to evaluate the expected likelihood surface when it is not gaussian in the parameter space, straightforwardly generalizing the DALI technique of Sellentin et al. (2014). Finally, the generalised Fisher Matrix will be implemented in the Fisher4Cast software, available at http://www.mathworks.com/matlabcentral/fileexchange/20008-fisher-matrix-toolbox-fisher4cast.

Acknowledgments
We are grateful to the organisers of the Cape Town International Cosmology School, where this work started as a student project, to Roberto Trotta, Daniel Mortlock and Andrew Jaffe for useful discussions, and to the anonymous referee for very helpful comments and suggestions.

APPENDIX A: PROOF THAT $\det \mathsf{C} \det \mathsf{A} = \det \mathsf{R}$

With $\det \mathsf{C} = \det(\mathsf{C}_{\rm XX}) \det(\mathsf{C}_{\rm YY} - \mathsf{C}_{\rm XY}^T \mathsf{C}_{\rm XX}^{-1} \mathsf{C}_{\rm XY}) = \det(\mathsf{C}_{\rm XX}) \det(\mathsf{E}^{-1})$, we have, reversing the order of the determinants,

$$\det \mathsf{C} \det \mathsf{A} = \det(\mathsf{E}^{-1}) \det(\mathsf{C}_{\rm XX}) \det \left[\mathsf{C}_{\rm XX}^{-1} + (\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} - \mathsf{T}^T)\mathsf{E}(\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} - \mathsf{T}^T)^T \right] \tag{A1}$$

Now, since $\det \mathsf{U} \det \mathsf{V} = \det(\mathsf{UV})$ for any square matrices,

$$\det \mathsf{C} \det \mathsf{A} = \det(\mathsf{E}^{-1}) \det \left[\mathsf{I} + (\mathsf{C}_{\rm XY} - \mathsf{C}_{\rm XX}\mathsf{T}^T)\mathsf{E}(\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} - \mathsf{T}^T)^T \right]. \tag{A2}$$

Now we use Sylvester's Determinant Theorem, $\det(\mathsf{I} + \mathsf{UV}) = \det(\mathsf{I} + \mathsf{VU})$ where we take $\mathsf{V} = \mathsf{E}(\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} - \mathsf{T}^T)^T$:

$$\det \mathsf{C} \det \mathsf{A} = \det(\mathsf{E}^{-1}) \det \left[\mathsf{I} + \mathsf{E}(\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} - \mathsf{T}^T)^T(\mathsf{C}_{\rm XY} - \mathsf{C}_{\rm XX}\mathsf{T}^T) \right]. \tag{A3}$$

Using $\det \mathsf{U} \det \mathsf{V} = \det(\mathsf{UV})$ again, and expanding E^{-1},

$$\begin{aligned}\det \mathsf{C} \det \mathsf{A} &= \det\left[\mathsf{E}^{-1} + (\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} - \mathsf{T}^T)^T(\mathsf{C}_{\rm XY} - \mathsf{C}_{\rm XX}\mathsf{T}^T)\right] \\ &= \det\left[\mathsf{C}_{\rm YY} - \mathsf{C}_{\rm XY}^T\mathsf{C}_{\rm XX}^{-1}\mathsf{C}_{\rm XY} + (\mathsf{C}_{\rm XY}^T\mathsf{C}_{\rm XX}^{-1} - \mathsf{T})(\mathsf{C}_{\rm XY} - \mathsf{C}_{\rm XX}\mathsf{T}^T)\right] \\ &= \det\left[\mathsf{C}_{\rm YY} - \mathsf{C}_{\rm XY}^T\mathsf{T}^T - \mathsf{T}\mathsf{C}_{\rm XY} + \mathsf{T}\mathsf{C}_{\rm XX}\mathsf{T}^T\right] = \det \mathsf{R}. \qquad \text{(A4)}\end{aligned}$$

APPENDIX B: GENERALISATION TO NON-UNIFORM PRIOR, OR PARENT DISTRIBUTION

We now generalise the method to apply to cases where the prior in $\mathbf{x}$ is not uniform. We illustrate this with a simplifying assumption that the prior is a gaussian of specified width, and demonstrate that in the limit of a prior width which is much larger than the errors in $\mathbf{x}$, we recover the results in the main text, and we expect this to hold for any broad prior. We can consider a prior which is dependent on each point, with a mean vector $\mathbf{a}$ and variance Σ (we assume that Σ is a diagonal matrix). In the normal case where the abscissa values are drawn from the same distribution, then all elements of $\mathbf{a}$ are identical, and Σ is proportional to the identity matrix.

Assuming a gaussian prior

$$p(\mathbf{x}) \propto \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T\Sigma^{-1}(\mathbf{x}-\mathbf{a})\right] \qquad \text{(B1)}$$

we get for the posterior

$$\mathcal{P} \propto \int \frac{1}{\sqrt{\det \mathsf{C}}} \exp\left\{-\frac{1}{2}\left[Q + (\mathbf{x}-\mathbf{a})^T\Sigma^{-1}(\mathbf{x}-\mathbf{a})\right]\right\} d^N\mathbf{x}\,, \qquad \text{(B2)}$$

Defining $\tilde{\mathbf{X}} \equiv \mathbf{X} - \mathbf{a}$, we get

$$Q + (\mathbf{x}-\mathbf{a})^T\Sigma^{-1}(\mathbf{x}-\mathbf{a}) = \tilde{\mathbf{x}}^T\tilde{\mathsf{A}}\tilde{\mathbf{x}} - 2\tilde{\mathbf{B}}^T\tilde{\mathbf{x}} + \tilde{\mathbf{X}}^T\Sigma^{-1}\tilde{\mathbf{X}} + Q' \qquad \text{(B3)}$$

where

$$\tilde{\mathsf{A}} = \mathsf{A} + \Sigma^{-1} \qquad \text{(B4)}$$

$$\tilde{\mathbf{B}}^T = \mathbf{B}^T + \tilde{\mathbf{X}}^T\Sigma^{-1}. \qquad \text{(B5)}$$

We perform the gaussian integral as before, finding

$$\mathcal{P} \propto \frac{1}{\sqrt{\det \mathsf{C} \det \tilde{\mathsf{A}}}} \exp\left[-\frac{1}{2}\left(-\tilde{\mathbf{B}}^T\tilde{\mathsf{A}}^{-1}\tilde{\mathbf{B}} + \tilde{\mathbf{X}}^T\Sigma^{-1}\tilde{\mathbf{X}} + Q'\right)\right]. \qquad \text{(B6)}$$

In the case when the prior in $\mathbf{x}$ is informative, then there is information in the values of $\mathbf{X}$, so the data vector should include both $\mathbf{X}$ and $\mathbf{Y}$. The likelihood is then

$$\mathcal{P} \propto \frac{1}{\sqrt{\det \mathsf{C} \det \tilde{\mathsf{A}}}} \exp\left(-\frac{1}{2}Q_{\rm YX}\right) \qquad \text{(B7)}$$

where

$$Q_{\rm YX} = (\tilde{\mathbf{Y}}, \tilde{\mathbf{X}})^T\mathsf{J}(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}) \qquad \text{(B8)}$$

and, collecting terms and using the Woodbury identity again, we find

$$\mathsf{J} = \begin{pmatrix} \mathsf{E} - (\mathsf{H}^T - \mathsf{E}\mathsf{T})(\mathsf{A} + \Sigma^{-1})^{-1}(\mathsf{H} - \mathsf{T}^T\mathsf{E}) & (\mathsf{H}^T - \mathsf{E}\mathsf{T})(\mathsf{A} + \Sigma^{-1})^{-1}\Sigma^{-1} \\ \Sigma^{-1}(\mathsf{A} + \Sigma^{-1})^{-1}(\mathsf{H} - \mathsf{T}^T\mathsf{E}) & (\Sigma + \mathsf{A}^{-1})^{-1} \end{pmatrix}. \qquad \text{(B9)}$$

In the limit of an infinitely broad prior, we see that, as expected, $\tilde{\mathbf{X}}$ contains no useful information, and the likelihood depends only on $\tilde{\mathbf{Y}}$, with the quadratic simplifying to $Q_{\rm YX} \to Q_{\rm Y} \equiv \tilde{\mathbf{Y}}^T\mathsf{R}\tilde{\mathbf{Y}}$, and as expected, we recover the results of the main text.

To investigate departures from the main text result, we consider terms linear in $\Sigma^{-1}\mathsf{A}^{-1}$. This approximation only makes sense if

$$\lim_{n\to\infty} \left(\Sigma^{-1}\mathsf{A}^{-1}\right)^n = 0\,. \qquad \text{(B10)}$$

As Σ is a diagonal matrix, the elements of the matrix $(\Sigma^{-1}\mathsf{A}^{-1})^n$ are given by

$$\begin{aligned}\left[\left(\Sigma^{-1}\mathsf{A}^{-1}\right)^n\right]_{ij} &= \left(\left[\Sigma^{-1}\right]_{ii}\left[\mathsf{A}^{-1}\right]_{ii}\right)^{n-1}\left[\Sigma^{-1}\right]_{ii}\left[\mathsf{A}^{-1}\right]_{ij} \\ &= \left(\left[\mathsf{A}^{-1}\right]_{ii}/\Sigma_{ii}\right)^{n-1}\left[\mathsf{A}^{-1}\right]_{ij}/\Sigma_{ii} \qquad \text{(B11)}\end{aligned}$$

Thus condition (B10) is fulfilled if

$$\left[\mathsf{A}^{-1}\right]_{ii} \ll \Sigma_{ii} \tag{B12}$$

for all i. We will assume this and neglect higher order terms in $\Sigma^{-1}\mathsf{A}^{-1}$. Then we can approximate $\tilde{\mathsf{A}}^{-1}$ by

$$\begin{aligned}\tilde{\mathsf{A}}^{-1} &= \left(\mathsf{A}+\Sigma^{-1}\right)^{-1} \\ &= \mathsf{A}^{-1}\left(\mathsf{I}+\Sigma^{-1}\mathsf{A}^{-1}\right)^{-1} \simeq \mathsf{A}^{-1}\left(\mathsf{I}-\Sigma^{-1}\mathsf{A}^{-1}\right)\end{aligned} \tag{B13}$$

Inserting this result in equation (B6), we get

$$\mathcal{P} \propto L_0 L_1 \tag{B14}$$

with

$$L_0 = \frac{1}{\sqrt{\det \mathsf{C}\det \mathsf{A}}}\exp\left(-\frac{1}{2}\tilde{\mathbf{Y}}^T\mathsf{R}^{-1}\tilde{\mathbf{Y}}\right) \tag{B15}$$

and

$$L_1 = \frac{1}{\sqrt{\det\left(\mathsf{I}+\Sigma^{-1}\mathsf{A}^{-1}\right)}}\exp\left[-\frac{1}{2}\left(\mathsf{A}^{-1}\mathbf{B}+\tilde{\mathbf{X}}\right)^T\Sigma^{-1}\left(\mathsf{A}^{-1}\mathbf{B}+\tilde{\mathbf{X}}\right)\right]. \tag{B16}$$

L_0 is the zeroth order result from the main text.

The Fisher matrix is then given by

$$\mathsf{F}_{\alpha\beta} = \mathsf{F}^{(0)}_{\alpha\beta} + \mathsf{F}^{(1)}_{\alpha\beta} \tag{B17}$$

with

$$\mathsf{F}^{(i)}_{\alpha\beta} = -\left\langle \frac{\partial^2 \ln L_i}{\partial\theta_\alpha \partial\theta_\beta}\right\rangle \qquad i=0,1. \tag{B18}$$

We already know the result for $\mathsf{F}^{(0)}_{\alpha\beta}$, so we just need to calculate the first-order term:

$$\mathsf{F}^{(1)}_{\alpha\beta} = \left\langle \frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\left[\frac{1}{2}\ln\det\left(\mathsf{I}+\Sigma^{-1}\mathsf{A}^{-1}\right)+\frac{1}{2}\left(\mathsf{A}^{-1}\mathbf{B}+\tilde{\mathbf{X}}\right)^T\Sigma^{-1}\left(\mathsf{A}^{-1}\mathbf{B}+\tilde{\mathbf{X}}\right)\right]\right\rangle. \tag{B19}$$

Using the approximation

$$\ln\det\left(\mathsf{I}+\Sigma^{-1}\mathsf{A}^{-1}\right) = \mathrm{Tr}\ln\left(\mathsf{I}+\Sigma^{-1}\mathsf{A}^{-1}\right) \simeq \mathrm{Tr}\left(\Sigma^{-1}\mathsf{A}^{-1}\right) \tag{B20}$$

and with $\langle\tilde{\mathbf{Y}}\rangle = 0$, $\langle\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\rangle = \mathsf{R}$, and $\tilde{\mathbf{Y}}_{,\alpha} = -\boldsymbol{\mu}_{,\alpha}$ we find after some tedious calculations

$$\begin{aligned}\mathsf{F}^{(1)}_{\alpha\beta} &= \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}\{\mathsf{A}^{-1}\}_{,\alpha\beta}+\left\{\left(\mathsf{H}^T-\mathsf{E}\mathsf{T}\right)\mathsf{A}^{-1}\Sigma^{-1}\mathsf{A}^{-1}\left(\mathsf{H}-\mathsf{T}^T\mathsf{E}\right)\right\}_{,\alpha\beta}\mathsf{R}\right] \\ &\quad -\tilde{\mathbf{X}}^T\Sigma^{-1}\left\{\mathsf{A}^{-1}\left(\mathsf{H}-\mathsf{T}^T\mathsf{E}\right)\boldsymbol{\mu}\right\}_{,\alpha\beta}+\tilde{\mathbf{X}}^T\Sigma^{-1} \\ &\quad \left\{\mathsf{A}^{-1}\left(\mathsf{H}-\mathsf{T}^T\mathsf{E}\right)\right\}_{,\alpha\beta}\boldsymbol{\mu}+\boldsymbol{\mu}^T_{,\alpha}\left(\mathsf{H}^T-\mathsf{E}\mathsf{T}\right)\mathsf{A}^{-1}\Sigma^{-1}\mathsf{A}^{-1}\left(\mathsf{H}-\mathsf{T}^T\mathsf{E}\right)\boldsymbol{\mu}_{,\beta}\,.\end{aligned} \tag{B21}$$

As $\{\mathsf{A}^{-1}\}_{,\alpha} = -\mathsf{A}^{-1}\mathsf{A}_{,\alpha}\mathsf{A}^{-1}$, each term in (B21) contains the factor $\Sigma^{-1}\mathsf{A}^{-1}$, so $\mathsf{F}^{(1)}_{\alpha\beta}$ gives the first-order corrections in terms of this parameter.

REFERENCES

Acquaviva V., Gawiser E., Bickerton S.J., Grogin N.A., Guo Y., Lee S.-K., 2012, The Astrophysical Journal, 749, 72

Albrecht A., Bernstein G., Cahn R., Freedman W.L., Hewitt J., Hu W., Huth J., Kamionkowski M., Kolb E.W., Knox L., Mather J.C., Staggs S., Suntzeff N.B., 2006, arXiV:0609591

Bassett B.A., Fantaye Y., Hlozek R., Kotze J., 2009, arXiv.org, astro-ph.CO

Coe D., 2009, Arxiv preprint arXiv:0906.4123

Dark Energy Survey Collaboration, 2005, arXiV:0510346

Cunha C., 2009, Physical Review D, 79, 63009

D'Agostini G., 2005, arXiV:0511182

Fisher R.A., 1935, J. Roy. Stat. Soc., 98, 39

Gull S.F., 1989, in Skilling J. (ed.), in "Maximum entropy and Bayesian methods", Kluwer publishing, 511, 518

Hogg D.W., Bovy J., Lang D., 2010, arXiV:1008.4686

Kelly B.C., 2011, in Feigelson E., Babu J. (eds.), "Statistical Challenges in Modern Astronomy V", Penn State, arXiV:1112.1745

Kim A.G., Miquel R., 2007, Astroparticle Physics, 28, 448

Kitching T.D., Heavens A. F., Verde L., Serra P., Melchiorri A., 2008, PRD, 77, 3008
Mandel K.S., Narayan G., Kirshner R.P., 2011, ApJ, 731, 120
March M.C., Trotta R., Berkes P., Starkman G.D., Vaudrevange P.M., 2011, MNRAS, 418, 2308
Refregier A., Amara A., Kitching T. D., Rassat A., 2011, A&A, 528, 33
Schlegel D. et al., 2011, arXiV:1106.1706
Sellentin E., Quartin M., Amendola L., 2014, arXiV:1401.6892
Taylor A., Heavens A., Ballinger B., Tegmark M., 1997, in "Proceedings of the Particle Physics and Early Universe Conference" (PPEUC), University of Cambridge, arXiv:9707265
Tegmark M., Taylor A., Heavens A., 1997, ApJ, 480, 22
Vogeley, M., Szalay A., 1996, ApJ, 465, 34
Wolz L. et al., 2012, JCAP, 9, 009
Woodbury M.A., 1950, Statistical Research Group, Memo Rep. No. 42