Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **408,** 865–875 (2010) doi:10.1111/j.1365-2966.2010.17201.x

Analytic methods for cosmological likelihoods

A. N. Taylor⋆ and T. D. Kitching⋆

Scottish Universities Physics Alliance (SUPA), Institute for Astronomy, School of Physics and Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ

Accepted 2010 June 15. Received 2010 May 14; in original form 2010 March 4

ABSTRACT

We present general, analytic methods for cosmological likelihood analysis and solve the 'many parameters' problem in cosmology. Maxima are found by Newton's method, while marginalization over nuisance parameters, and parameter errors and covariances are estimated by analytic marginalization of an arbitrary likelihood function, expanding the log-likelihood to second order, with flat or Gaussian priors. We show that information about remaining parameters is preserved by marginalization. Marginalizing over all parameters, we find an analytic expression for the Bayesian evidence for model selection. We apply these methods to data described by Gaussian likelihoods with parameters in the mean and covariance. These methods can speed up conventional likelihood analysis by orders of magnitude when combined with Markov chain Monte Carlo methods, while Bayesian model selection becomes effectively instantaneous.

Key words: methods: analytical – methods: data analysis – methods: statistical – cosmology: theory – large-scale structure of Universe.

1 INTRODUCTION

There is now a standard model of cosmology, Λ cold dark matter (ΛCDM), which has substantial predictive power but is highly unsatisfactory from a theoretical viewpoint. The most serious of these is the unknown nature of the dominant dark energy component driving the accelerated expansion of the Universe. This may be due to a new force of nature, or possibly a break-down of Einstein gravity on large scales. Without a clear direction of how to progress beyond a phenomenological picture to a more fundamental theory, attention is turning to proposing a wide range of modified or alternative models to the standard model and use observations as a guide to future progress.

To realize this a number of large and challenging observational programmes are being planned and carried out, e.g. ESA's Planck Cosmic Microwave Background mission, the Canada–France–Hawaii Telescope Legacy Survey (CFHTLS), ESA's Visible and Infrared Survey Telescope for Astronomy (VISTA) and VLT Survey Telescope (VST), the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), the Dark Energy Survey (DES), the Large Synoptic Survey Telescope (LSST), ESA's proposed Euclid satellite, the NASA/DOE proposed Joint Dark Energy Mission (JDEM) and the Square-Kilometre Array (SKA). One of the main aims of these large data sets is to distinguish between diverse competing models, some with large parameter spaces. The ΛCDM model, and basic extensions, contains some 18 parameters, ($\Omega_{\rm m}$, $\Omega_{\rm b}$, $\Omega_{\rm de}$, Ω_ν, w_0, $w_{\rm a}$, h, $A_{\rm s}$, $n_{\rm s}$, $\alpha_{\rm s}$, A_T, n_T, τ, b, $f_{\rm NL}$, $A_{\rm iso}$, γ, η), covering the dark matter, dark energy, initial conditions and gravity sectors. Such large parameter spaces become a problem to investigate, while fundamental models of dark energy or modified gravity may have many more parameters which are not well described by these phenomenological parameters.

The analysis of these large-scale data sets is not limited by shot noise, data volume or the volume of the Universe covered. The main limitation is our ability to understand and remove, to high accuracy, systematic effects in the data. For example, we may not precisely know the calibration factor, beam size and shape or effect of Galactic foreground contamination in cosmic microwave background (CMB) experiments; the calibration and effect of outliers in photometric redshift surveys; scale-dependent and stochastic bias in galaxy redshift surveys; calibration of cosmic shear or intrinsic alignment effects in weak lensing surveys; or environmental effects and evolution in Type Ia supernovae. These systematic effects are generally parametrized by a set of nuisance parameters, which themselves must to be constrained by data. The number of these nuisance parameters can vastly outweigh the number of cosmological parameters. The size of these large parameter spaces for a likelihood analysis is the 'many parameters' problem.

We also need a systematic approach to discriminating between what is becoming a large number of competing cosmological models for dark energy and modified gravity. The Bayesian approach to model selection is to evaluate the *evidence*, the probability of model given the data, for all possible cosmological and nuisance parameter space. For a large number models, each with a large number of cosmological and nuisance parameters, this can be an immense task.

⋆E-mail: ant@roe.ac.uk (ANT);tdk@roe.ac.uk (TDK)

The standard approach to the analysis of cosmological data sets is through a likelihood analysis of the model parameter space (e.g. Kaiser 1988; Heavens & Taylor 1995; Verde et al. 2003). Parameter values are given by the maximum, or mean, of the likelihood function, and parameter errors and covariances are given by the shape of the marginalized likelihood surface around the maximum. Since we are not directly interested in nuisance parameters which characterize systematic effects, these are marginalized out. To evaluate the Bayesian evidence we marginalize over the entire parameter space, both cosmological and nuisance to find the probability of the model.

The likelihood surface can be mapped out numerically using Markov chain Monte Carlo (MCMC) methods (Gamerman 1997; Lewis & Bridle 2002; MacKay 2003), where the likelihood distribution is sampled by a cloud of points whose density follows the likelihood. Marginalization is then carried out by projecting the points on to subsets of the parameter space. As efficient as this is, when the number of parameters and nuisance parameters becomes large, or even infinite, this become unfeasible. MCMC is not an efficient or accurate way to find the maximum of the likelihood, and mean values are often quoted. The MCMC method can also be sensitive to the choice of priors, and insensitive to sharply peaked and strongly degenerate likelihood surfaces. Methods have evolved to compensate for this, including using physical parameters (Kosowsky, Milosavljevic & Jimenez 2002) or rotating to orthogonal parameter sets (Tegmark et al. 2004). However, the effect of priors on these spaces is less transparent.

An alternative approach to numerical marginalization is to approximate the likelihood in parameter space as a Gaussian and analytically marginalize (Bretthorst 1988; Gull 1989; Bridle et al. 2002; MacKay 2003). Bridle et al. (2002) apply this method in cosmology to marginalize over nuisance parameters appearing in the mean of a Gaussian likelihood. This approach is exact when the parameters are Gaussian distributed such as the amplitude of the mean, and this is publicly available in CosmoMC[1] (Lewis & Bridle 2002). An analytic marginalization method has also been developed for evaluating the Bayesian evidence, using the saddle point, or Laplace, approximation to marginalize over all parameters around the peak of the likelihood (e.g. MacKay 2003; Trotta 2008). However, this does not evaluate the absolute evidence. There is no general treatment of analytic marginalization which will accommodate both of these, and even more general, situations. In this paper we present a new, self-consistent and general framework in which to maximize and marginalize over an arbitrary likelihood function, to remove nuisance parameters, estimate marginalized projections of parameter space and derive an analytic expression for the Bayesian evidence.

The paper is set out as follows. In Section 2 we describe likelihood methods for parameter estimation and set out the general approach for maximization and marginalization over nuisance parameters for an arbitrary likelihood function with flat or Gaussian priors. We show that the marginalized likelihood function preserves information on cosmological parameters. In Section 3 we show how to apply the method to the specific case of a multivariate Gaussian-distributed data where the cosmological and systematic information is contained in the mean and covariance. In Section 4 we present some applications: marginalization over an amplitude, projections of parameter space and semi-analytic marginalization. We show how our methods can applied to find a solution to the problem of Bayesian evidence in Section 5, and discuss some aspects of model selection in model-space. Finally, in Section 6 we present our conclusions.

[1]http://cosmologist.info

2 ANALYTIC LIKELIHOOD ANALYSIS

Assuming a model, $\mathcal{M}$, for a cosmological data set, $\boldsymbol{D}$, which is parametrized by a set of $N_{\rm p}$ parameters, θ, the conditional probability distribution of the data is given by the likelihood function, $L = p(\boldsymbol{D}|\boldsymbol{\theta}, \mathcal{M})$. We can transform from the likelihood function to the posteriori probability for the parameters given the data, $p(\boldsymbol{\theta}|\boldsymbol{D}, \mathcal{M})$, using Bayes' theorem;

$$p(\boldsymbol{\theta}|\boldsymbol{D}, \mathcal{M}) = \frac{L(\boldsymbol{D}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{D}|\mathcal{M})}, \tag{1}$$

where $p(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution of the parameters assumed before the analysis. The normalizing distribution, $p(\boldsymbol{D}|\mathcal{M})$, is called the *evidence*. Priors are commonly assumed to be either flat, where the distribution is a top-hat with constant value over some parameter range and zero outside, or Gaussian with a mean constrained by earlier experiments. The posterior distribution is then maximized with respect to the $N_{\rm p}$ cosmological parameters in the model. Marginalization of the posteriori or likelihood function is required if we have a subset of M parameters, ψ, which we want to integrate over,

$$p(\boldsymbol{\theta}|\mathcal{M}) = \int \mathrm{d}^M\psi \; p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathcal{M}). \tag{2}$$

The $\boldsymbol{\psi}$ parameters may be nuisance parameters which characterize some systematic effect, or some of the cosmological parameters, $\boldsymbol{\theta}$, whose effect we want to integrate over when we do not have an accurate understanding of the effect (e.g. the normalization of galaxy perturbations due to galaxy bias). We may also want to project out the likelihood surface to lower dimensions to study the distribution, or even marginalize over all of the $N_{\rm p} + M$ nuisance and cosmological parameters if we want to estimate the evidence.

Now consider an arbitrary likelihood function, $L(\boldsymbol{D}|\boldsymbol{\Phi}, \mathcal{M})$, which depends on a set of cosmological parameters, $\boldsymbol{\theta}$, and on a set of marginalization parameters, $\boldsymbol{\psi}$, which we want to integrate over, where we have combined all parameters into $\boldsymbol{\Phi} = (\boldsymbol{\theta}, \boldsymbol{\psi})$. We begin by defining the log-likelihood, $\mathcal{L}$, of the likelihood function

$$\mathcal{L} = -2 \ln L. \tag{3}$$

This can be expanded around an arbitrary point, $\boldsymbol{\Phi}_0$, in the full parameter space to second-order

$$\mathcal{L} = \mathcal{L}_0 + \delta\Phi_\mu \mathcal{L}_\mu + \frac{1}{2}\delta\Phi_\mu \delta\Phi_\nu \mathcal{L}_{\mu\nu}, \tag{4}$$

where $\mathcal{L}_\nu = \partial_\nu \mathcal{L}_0$ and $\mathcal{L}_{\nu\mu} = \partial_\nu \partial_\mu \mathcal{L}_0$ are evaluated at $\boldsymbol{\Phi}_0$, and where we denote derivatives with respect to a nuisance parameter by Greek indices.

2.1 Maximizing the likelihood

We first want to find the minimum of the log-likelihood function in the full $N_{\rm P} + M$ cosmological and nuisance parameter space. Differentiating equation (4) with respect to the parameters and setting the gradient to zero, we find the displacement between the fiducial point and the peak of the likelihood as

$$\delta\Phi_\mu = -\mathcal{L}_\nu \mathcal{L}^{-1}_{\nu\mu}. \tag{5}$$

If the likelihood is close to Gaussian, we can find the maximum of the likelihood in a single step. If the likelihood is non-Gaussian, but

smooth, we can iterate towards the peak. This is Newton's method for finding the peak of the likelihood (e.g. Press et al. 1989). The assumption of Gaussianity in equation (4) is therefore benign since we can maximize the likelihood for non-Gaussian distributions.

2.2 Analytic marginalization

We now want to marginalize over the $\boldsymbol{\psi}$ nuisance parameters. Locally expanding the likelihood in the $\boldsymbol{\psi}$ parameters yields

$$\mathcal{L} = \mathcal{L}_0 + \delta\psi_\alpha \mathcal{L}_\alpha + \frac{1}{2}\delta\psi_\alpha \delta\psi_\beta \mathcal{L}_{\alpha\beta}, \tag{6}$$

where the indices α and β refer to nuisance parameters. Analytically marginalizing over $\boldsymbol{\psi}$ (see Appendix A for details), assuming a non-zero flat prior in the volume V_ψ of $\boldsymbol{\psi}$-space, $p(\boldsymbol{\psi}|\mathcal{M}) = 1/V_\psi$, yields

$$\mathcal{L} = \mathcal{L}_0 - \frac{1}{2}\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta + \mathrm{Tr}\,\ln\left(V_\psi^{2/M}\mathcal{L}_{\alpha\beta}\right), \tag{7}$$

where we have dropped an unimportant constant of $-M\,\ln(4\pi)$. This is the marginalized log-likelihood function under the assumption of local Gaussianity in these directions. We consider methods for dealing with non-Gaussianity in Section 4.3.

The first term, $\mathcal{L}_0 = \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\psi} = \boldsymbol{\psi}_0)$ is the conditional likelihood at fixed ψ. The second term in equation (7), which is quadratic in $\mathcal{L}_\alpha$, has an intuitive meaning. Although we have fixed the values of $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ at their maximum in the full parameter space, and where the gradient is zero, the likelihood is still a function of the remaining parameters, $\boldsymbol{\theta}$. As we move in parameter space away from the maximum along one of the directions of $\boldsymbol{\theta}$, the peak will move away from $\boldsymbol{\psi}_0$, unless the parameters are uncorrelated, and the gradient $\mathcal{L}_\alpha$ will be non-zero. This term then describes the full shape of the likelihood and the coupling between the marginalized parameters and the remaining parameters. Its presence removes the dependence of the likelihood on the marginalized parameters, and widens the distribution.

The third, well-known, term accounts for the volume of marginalized parameter space with significant likelihood, and is called the *Occam factor*. The presence of the curvature of the log-likelihood, through $\mathcal{L}_{\alpha\beta}$, shows that this expression is sensitive to information in the data itself about the systematic nuisance parameters. Note that we have made no assumptions about the form of the likelihood function in $\boldsymbol{\theta}$ space, only that we can approximate the peak of the likelihood function in the marginalized $\boldsymbol{\psi}$ parameter space by a multivariate Gaussian. Analytic marginalization does not suffer from prior boundary problems, since the full likelihood space is marginalized over, and infinitely resolves the peak of the likelihood.

We can derive the marginalized likelihood in a second, more illuminating, way. We can use the expansion given by equation (6) to find the displacement of a fixed point in nuisance parameter space from the peak of the likelihood,

$$\delta\psi_\alpha = -\mathcal{L}_\beta \mathcal{L}^{-1}_{\alpha\beta}. \tag{8}$$

Substituting this back into equation (6) we find that maximum value of the likelihood is

$$\mathcal{L}_{\max} = \mathcal{L}_0 - \frac{1}{2}\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta. \tag{9}$$

The first two terms in equation (7) are just the maximum likelihood value, while the third term is just the width of the likelihood curve. This shows us that the marginalized likelihood is independent of the choice of $\boldsymbol{\psi}_0$, when $\mathcal{L}(\boldsymbol{\psi})$ is Gaussian, since the second term in equation (9) corrects the likelihood estimated at $\boldsymbol{\psi}_0$ to the value at the peak. In Appendix B, we derive the mean and variance of the likelihood from its Generating Function.

Analytic marginalization preserves information about cosmological parameters. Expanding equation (7) to lowest order in the remaining cosmological parameters, $\Delta\boldsymbol{\theta}$, around the peak of the ensemble averaged likelihood keeping the curvature $\mathcal{L}_{\alpha\beta}$ fixed at its expectation value, we find

$$\mathcal{L} = \mathcal{L}_0 + \Delta\theta_i \Delta\theta_j \left[\langle\mathcal{L}_{ij}\rangle - \langle\mathcal{L}_{i\alpha}\rangle\langle\mathcal{L}_{\alpha\beta}\rangle^{-1}\langle\mathcal{L}_{\beta j}\rangle\right], \tag{10}$$

where Arabic indices i and j indicate cosmological parameters. Here we can identify the Schur complement (e.g. Zhang 2005) of the marginalized Fisher information matrix for cosmological parameters,

$$F^M_{ij} = F_{ij} - F_{i\alpha}F^{-1}_{\alpha\beta}F_{\beta j}, \tag{11}$$

where

$$F_{\mu\nu} = \frac{1}{2}\langle\mathcal{L}_{\mu\nu}\rangle \tag{12}$$

is the full $(N_\mathrm{p} + M)$-dimensional Fisher matrix (see, e.g. Tegmark, Taylor & Heavens 1997) for cosmological parameters and systematic nuisance parameters. The indices (μ, ν) extend over all (i, j) and (α, β). Equation (11) is identical to the Fisher matrix found by maximizing the pre-marginalized likelihood and then marginalizing over the nuisance parameters. Hence, at the level of Fisher matrices, no information is lost by analytic marginalization.

When we have a Gaussian prior on the nuisance parameters, the log-likelihood becomes

$$\mathcal{L} = \mathcal{L}_0 + \delta\psi_\alpha \mathcal{L}_\alpha + \frac{1}{2}\delta\psi_\alpha[\mathcal{L}_{\alpha\beta} + 2C^{-1}_{\alpha\beta}]\delta\psi_\beta + \mathrm{Tr}\,\ln C_{\alpha\beta}, \tag{13}$$

where $C_{\alpha\beta}$ is the prior covariance matrix. The maximum is now found at

$$\delta\psi_\alpha = -\mathcal{L}_\beta[\mathcal{L}_{\alpha\beta} + 2C^{-1}_{\alpha\beta}]^{-1}, \tag{14}$$

while marginalization leads to

$$\mathcal{L} = \mathcal{L}_0 - \frac{1}{2}\mathcal{L}_\alpha[\mathcal{L}_{\alpha\beta} + 2C^{-1}_{\alpha\beta}]^{-1}\mathcal{L}_\beta + \mathrm{Tr}\,\ln\left(\delta^K_{\alpha\beta} + \frac{1}{2}C_{\alpha\delta}\mathcal{L}_{\delta\beta}\right). \tag{15}$$

3 GAUSSIAN LIKELIHOODS

Let us assume the statistical properties of the data, $\boldsymbol{D}$, can be modelled by a multivariate Gaussian distribution, $L(\boldsymbol{D}|\boldsymbol{\theta}, \boldsymbol{\psi})$ which depends only on a mean, $\boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \langle\boldsymbol{D}\rangle$, and a covariance matrix, $\boldsymbol{C}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \langle\Delta\boldsymbol{D}\Delta\boldsymbol{D}^t\rangle$, where $\Delta\boldsymbol{D} = \boldsymbol{D} - \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the variation of the data about the mean. By definition $\langle\Delta\boldsymbol{D}\rangle = 0$. The Gaussian log-likelihood function is given by

$$\mathcal{L}_0 = \Delta\boldsymbol{D}\boldsymbol{C}^{-1}\Delta\boldsymbol{D}^t + \mathrm{Tr}\,\ln\boldsymbol{C}. \tag{16}$$

The cosmological and nuisance parameters can appear in both the mean of the data values or in the covariance. We consider each in turn, starting with parameters in the mean.

3.1 Parameters in the mean

If the nuisance parameters are in the mean, $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\psi})$, and we assume a flat prior on marginalization parameters, the gradient and curvature of the log-likelihood in parameter space is

$$\mathcal{L}_\alpha = -2\Delta\boldsymbol{D}^t\boldsymbol{C}^{-1}\boldsymbol{\mu}_\alpha, \tag{17}$$

$$\mathcal{L}_{\alpha\beta} = 2\left(\boldsymbol{\mu}_{\alpha}\boldsymbol{C}^{-1}\boldsymbol{\mu}_{\beta}^{t} - \Delta\boldsymbol{D}^{t}\boldsymbol{C}^{-1}\boldsymbol{\mu}_{\alpha\beta}\right). \quad (18)$$

The expectation value of the slope is $\langle\mathcal{L}_{\alpha}\rangle = 0$, while the expectation value of the curvature around the peak in parameter space is

$$\langle\mathcal{L}_{\alpha\beta}\rangle = 2F_{\alpha\beta} = 2\boldsymbol{\mu}_{\alpha}\boldsymbol{C}^{-1}\boldsymbol{\mu}_{\beta}^{t}. \quad (19)$$

If we choose to use the Fisher Matrix for the local curvature, the maximum of the Gaussian likelihood function lies at

$$\Phi_{\nu}^{\max} = \Phi_{\nu}^{0} + F_{\mu\nu}^{-1}\Delta\boldsymbol{D}^{t}\boldsymbol{C}^{-1}\boldsymbol{\mu}_{\mu}, \quad (20)$$

where $\boldsymbol{\Phi}^{0}$ is an arbitrary point in parameter space. Since the curvature is approximated by the Fisher matrix, this is a quasi-Newtonian method. Again if the likelihood is Gaussian in parameter space, this is exact and if not some iteration is required.

Marginalizing over the nuisance parameters assuming a flat prior, we find the likelihood function is again a Gaussian,

$$\mathcal{L} = \Delta\boldsymbol{D}\boldsymbol{C}_{\mathrm{M}}^{-1}\Delta\boldsymbol{D}^{t} + \mathrm{Tr}\,\ln V_{\psi}^{2/M}F_{\alpha\beta}, \quad (21)$$

where the marginalized data covariance matrix, $\boldsymbol{C}_{\mathrm{M}}$ is given by

$$\boldsymbol{C}_{\mathrm{M}} = \langle\Delta\boldsymbol{D}\Delta\boldsymbol{D}^{t}\rangle_{\mathrm{M}} = \left(\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{\mu}_{\alpha}^{t}F_{\alpha\beta}^{-1}\boldsymbol{\mu}_{\beta}\boldsymbol{C}^{-1}\right)^{-1}. \quad (22)$$

If we assume the curvature is given by its expectation value, the constant term, $\ln\det V_{\psi}^{2}F_{\alpha\beta}$ in equation (21), can be dropped and we can identify $\mathcal{L}$ with the χ^{2}-statistic and all our results still hold. Note that in these expressions, the parameter-dependence only appears in the mean in $\Delta\boldsymbol{D} = \boldsymbol{D} - \boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\psi}_{0})$. Everything else is fixed at the fiducial values, $\boldsymbol{\theta}_{0}$ and $\boldsymbol{\psi}_{0}$. We can also see from this solution that there is a requirement on the marginalized covariance matrix that it is positive definite, $\Delta\boldsymbol{D}\boldsymbol{C}_{\mathrm{M}}^{-1}\Delta\boldsymbol{D}^{t} > 0$, in order that the likelihood function has a maximum bound, however this is always true.

If we assume a Gaussian prior on the nuisance parameters, the marginalized data covariance matrix is regularized and can be simplified using the Woodbury matrix identity (Woodbury 1950) so that

$$\boldsymbol{C}_{\mathrm{M}} = \left(\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{\mu}_{\alpha}^{t}[F_{\alpha\beta} + C_{\alpha\beta}^{-1}]^{-1}\boldsymbol{\mu}_{\beta}\boldsymbol{C}^{-1}\right)^{-1} \quad (23)$$

$$= \boldsymbol{C} + C_{\alpha\beta}\boldsymbol{\mu}_{\alpha}\boldsymbol{\mu}_{\beta}^{t}, \quad (24)$$

where the last expression is explicitly positive-definite. Equations (23) and (24) have previously been derived by Bridle et al. (2002) using a somewhat different method for marginalizing over a Gaussian likelihood with a Gaussian prior and nuisance parameters in the mean. If we include a prior on nuisance parameters, the log-likelihood function becomes

$$\mathcal{L} = \Delta\boldsymbol{D}\boldsymbol{C}_{\mathrm{M}}^{-1}\Delta\boldsymbol{D}^{t} + \mathrm{Tr}\,\ln\boldsymbol{C}_{\mathrm{M}}, \quad (25)$$

again up to an unimportant normalization constant. We note that even if the cosmological parameters do not affect the covariance, the marginalized covariance, $\boldsymbol{C}_{\mathrm{M}}$, will gain a dependence on cosmological parameters through the mean.

3.2 Parameters in the covariance

If the parameters are in the data covariance matrix, $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{\theta}, \boldsymbol{\psi})$, the derivatives of the log-likelihood are

$$\mathcal{L}_{\alpha} = -\mathrm{Tr}\left(\partial_{\alpha}\ln\boldsymbol{C}\;\Delta\ln\boldsymbol{C}\right), \quad (26)$$

$$\mathcal{L}_{\alpha\beta} = \mathrm{Tr}\,[(\partial_{\alpha}\ln\boldsymbol{C})(\partial_{\beta}\ln\boldsymbol{C})(\boldsymbol{I} + 2\Delta\ln\boldsymbol{C}) - \boldsymbol{C}^{-1}(\partial_{\alpha}\partial_{\beta}\boldsymbol{C})\Delta\ln\boldsymbol{C}], \quad (27)$$

where $\partial_{\alpha}\ln\boldsymbol{C} = \boldsymbol{C}^{-1}\partial_{\alpha}\boldsymbol{C}$, $\Delta\ln\boldsymbol{C} = \Delta\boldsymbol{D}\boldsymbol{C}^{-1}\Delta\boldsymbol{D}^{t} - \boldsymbol{I}$ and $\langle\Delta\ln\boldsymbol{C}\rangle = 0$. The expectation values of the gradient is $\langle\mathcal{L}_{\alpha}\rangle = 0$ while the expectation of the curvature is given by

$$\langle\mathcal{L}_{\alpha\beta}\rangle = 2F_{\alpha\beta} = \mathrm{Tr}\,[(\partial_{\alpha}\ln\boldsymbol{C})(\partial_{\beta}\ln\boldsymbol{C})]. \quad (28)$$

If we assume the curvature is given by its expectation value, we find the peak is at

$$\delta\Phi_{\nu} = \frac{1}{2}F_{\nu\mu}^{-1}\,\mathrm{Tr}\left(\partial_{\mu}\ln\boldsymbol{C}\;\Delta\ln\boldsymbol{C}\right), \quad (29)$$

from the fiducial point in $\boldsymbol{\Phi}$ space. For a single-step estimate of the peak, this is equivalent to Tegmark's (1997) quadratic estimator. The analytically marginalized log-likelihood is

$$\mathcal{L} = \mathcal{L}_{0} - \frac{1}{4}\mathcal{L}_{\alpha}F_{\alpha\beta}^{-1}\mathcal{L}_{\beta} + \mathrm{Tr}\,\ln V_{\psi}^{2/M}F_{\alpha\beta}, \quad (30)$$

where $\mathcal{L}_{\alpha}$ is given by equation (26). To change the prior to a Gaussian, we again make the substitution

$$\mathcal{L} = \mathcal{L}_{0} - \frac{1}{4}\mathcal{L}_{\alpha}[F_{\alpha\beta} + C_{\alpha\beta}^{-1}]^{-1}\mathcal{L}_{\beta} + \mathrm{Tr}\,\ln\left(\delta_{\alpha\beta}^{K} + C_{\alpha\beta}F_{\alpha\beta}\right), \quad (31)$$

Again, we require that $\mathcal{L} > 0$ to bound the likelihood function.

4 APPLICATIONS

Having calculated the marginalized likelihoods for Gaussian-distributed data with parameters in both mean and covariance matrix, we now turn to two examples: marginalization over nuisance parameters and projections of the likelihood function in parameter space.

4.1 Systematic nuisance parameters

A simple, and well-known, example of a nuisance parameter is the normalization of the mean with a flat prior. This is an interesting case since the analysis is exact. Let the mean be given by $\boldsymbol{\mu} = A\boldsymbol{\mu}_{0}$, where the Fisher matrix for the amplitude, A, found from the data is given by $F_{AA} = (1/A^{2})\,\mathrm{Tr}\,[\boldsymbol{\mu}\boldsymbol{C}^{-1}\boldsymbol{\mu}^{t}]$, then

$$\boldsymbol{C}_{\mathrm{M}} = \left(\boldsymbol{C}^{-1} - \frac{\boldsymbol{C}^{-1}\boldsymbol{\mu}^{t}\boldsymbol{\mu}\boldsymbol{C}^{-1}}{\mathrm{Tr}\,[\boldsymbol{\mu}\boldsymbol{C}^{-1}\boldsymbol{\mu}^{t}]}\right)^{-1} \quad (32)$$

and the peak is found from equation (20). If we assume the covariance is diagonal, $C_{ij} = \sigma_{i}^{2}\delta_{ij}^{K}$, then the log-likelihood becomes

$$\mathcal{L} = \sum_{i}\frac{\Delta D_{i}^{2}}{\sigma_{i}^{2}} - \left(\frac{1}{\sum_{k}\mu_{k}^{2}/\sigma_{k}^{2}}\right)\left(\sum_{i}\frac{\Delta D_{i}\mu_{i}}{\sigma_{i}^{2}}\right)^{2}. \quad (33)$$

If we assume further that the mean values are Gaussian-distributed power spectra, $\mu_{k} = P_{k}$, their variance is given by $\sigma_{k}^{2} = 2\,P_{k}^{2}$, and the log-likelihood is

$$\mathcal{L} = \frac{1}{2}\sum_{\boldsymbol{k}}\left(\Delta\ln P_{k} - \overline{\Delta\ln P_{k}}\right)^{2}. \quad (34)$$

In the last expression $\Delta\ln P_{k} = [\widehat{P}_{k} - P_{k}(\boldsymbol{\theta})]/P_{k}$, where $\widehat{P}_{k}$ is the measured power, $\overline{x} = (1/N_{\mathrm{D}})\sum_{k}x_{k}$ and N_{D} is the number of data points. Hence the log-likelihood is positive-definite, and minimizing $\mathcal{L}$ is equivalent to minimizing the variance of $\Delta\ln P_{k}$. This expression makes sense as the second term removes any dependence on the best estimate of the calibration off-set from the likelihood. Equation (34) has an immediate cosmological application for removing the dependence of a linear galaxy bias on parameters estimated from the galaxy power spectrum, assuming the power spectrum passbands are independent.

More generally, we find the marginalized likelihood for multiple parameters is given by

$$\mathcal{L} = \frac{1}{2}\left[\sum_{\boldsymbol{k}} |\Delta \ln P_k|^2 - \frac{1}{2}\mathcal{L}_\alpha F^{-1}_{\alpha\beta}\mathcal{L}_\beta\right], \tag{35}$$

where the Fisher matrix and gradient of the log-likelihood are

$$F_{\alpha\beta} = \frac{1}{2}\sum_{\boldsymbol{k}} (\partial_\alpha \ln P_k)(\partial_\beta \ln P_k), \tag{36}$$

$$\mathcal{L}_\alpha = -\sum_{\boldsymbol{k}} \Delta \ln P_k\, \partial_\alpha \ln P_k, \tag{37}$$

and the peak of the likelihood is at

$$\delta\Phi_\mu = -\frac{1}{2}F^{-1}_{\mu\nu}\mathcal{L}_\nu. \tag{38}$$

If we want to include noise in these expressions, we can do so by substituting $P_k \rightarrow \mathrm{P_k} + \mathrm{N(r)}$, where $N(r)$ is the noise power, which may depend on position within the survey. For example in galaxy redshift surveys, $N(r) = 1/\bar{n}(r)$, and we should extend the summation over k to $\mathrm{Tr} \rightarrow \sum_k \int \mathrm{d}^3 r$. In the continuum limit, we would substitute $\mathrm{Tr} = \int \mathrm{d}^3 k/(2\pi)^3$ (see, e.g., Taylor & Watts 2001). For CMB or weak lensing analysis on the sky, we should substitute $P_k \rightarrow C_\ell$ and $\mathrm{Tr} \rightarrow \sum_\ell (2\ell+1)$, where we have implicitly assumed statistical isotropy and summed over the $2\ell + 1$ azimuthal modes. Finally, for 3D Cosmic Shear (e.g. Heavens, Kitching & Taylor 2006), where the covariance matrix is $\boldsymbol{C} = C_\ell^{\gamma\gamma}(z, z')$ we substitute $\mathrm{Tr} \rightarrow \sum_\ell (2\ell + 1) \int \mathrm{d}z \mathrm{d}z'$.

If the parameter appear in the covariance matrix, and the data have a Gaussian distribution, the log-likelihood distribution is given by

$$\mathcal{L}_0 = \mathrm{Tr}\,(\widehat{\boldsymbol{C}}\boldsymbol{C}^{-1} + \ln \boldsymbol{C}) = \mathrm{Tr}\,(\Delta \ln \boldsymbol{C} + \ln \boldsymbol{C}) + N_\mathrm{D}, \tag{39}$$

where $\widehat{\boldsymbol{C}} = \Delta \boldsymbol{D} \Delta \boldsymbol{D}^t$ and N_D is the number of data-points used. If again we use the example of marginalization over the normalization of the covariance matrix, $\boldsymbol{C} = A\, \boldsymbol{C}_0$, where the Fisher matrix is $F_{AA} = N_\mathrm{D}/2A^2$, the marginalized likelihood is

$$\mathcal{L} = \mathrm{Tr}\,(\Delta \ln \boldsymbol{C} + \ln \boldsymbol{C}) - \frac{1}{N_\mathrm{D}}\mathrm{Tr}\,[\Delta \ln \boldsymbol{C}\, \Delta \ln \boldsymbol{C}] + N_\mathrm{D}. \tag{40}$$

For a diagonal covariance matrix, the marginalized log-likelihood with parameters in the covariance can be written

$$\mathcal{L} = \sum_{\boldsymbol{k}} \left(\frac{\widehat{P}_k}{P_k} + \ln P_k\right) - \frac{1}{4}\mathcal{L}_\alpha F^{-1}_{\alpha\beta}\mathcal{L}_\beta. \tag{41}$$

Despite the different form of the term $\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta$ when the parameters appear in the data covariance matrix, in this limit this term is the same as when the parameters appear only in the mean (cf. equation 35).

4.1.1 Galaxy clustering

In Fig. 1 we show the likelihood, $L(\Omega_\mathrm{m}, A)$, for a joint measurement of the matter density parameter, Ω_m, and galaxy clustering amplitude, $A = b\sigma_8$, from the galaxy power spectrum, $P_\mathrm{g}(k)$. Here b is a linear bias parameter and σ_8 the variance of matter clustering in spheres of $8\,h^{-1}$ Mpc. The matter power spectrum is generated using the Eisenstein & Hu (1997) parametrization with a Smith et al. (2003) non-linear correction, and we have ignored the effect of redshift-space distortions.. We have assumed a fixed Hubble parameter, hence Ω_m determines the linear break scale in the matter power spectrum and amplitude of non-linear corrections. We assume a fiducial model with $\Omega_\mathrm{m} = 0.3$ and $b\sigma_8 = 1$. The error on the measured power is assumed to be sample-dominated, with negligible shot noise, given by $\sigma(k) = 2\pi P(k)/\sqrt{V k^3 \mathrm{d}\ln k}$ (e.g. Tegmark 1997), where we have assumed $V = 19.7\,h^{-3}\,\mathrm{Gpc}^3$ and spectroscopic redshifts and no redshift-space distortion. We include a wavenumber range up to $k_\mathrm{max} = 100\,h\,\mathrm{Mpc}^{-1}$. In the lower panel we show the two-parameter distribution and how Newton's method converges to the maximum likelihood. It is clear that after approximately three to four iterations the maximum likelihood is covered, even in this case of a highly non-Gaussian likelihood surface.

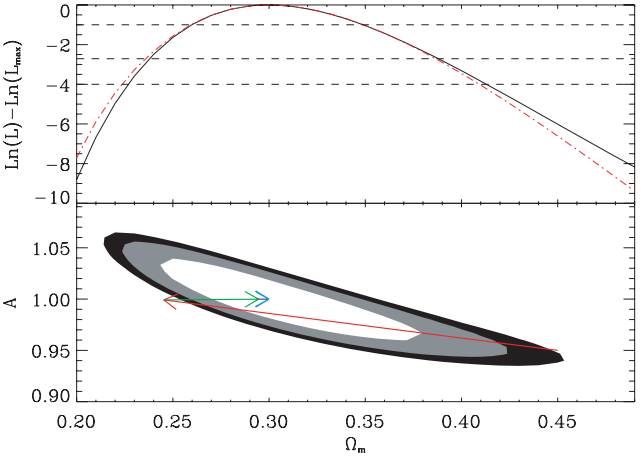

Figure 1. Example of marginalization over a nuisance parameter. The lower panel shows the two-parameter 1σ (68.3 per cent), 2σ (90 per cent) and 3σ (99.9 per cent) contours in white, grey and black for the matter density parameter, Ω_m, and a nuisance power-spectrum normalization parameter, $A = b\sigma_8$, for a measurement of the matter power spectrum for a survey covering an effective volume of $19.7\,h^{-3}\,\mathrm{Gpc}^3$ with negligible shot noise. The solid lines show the convergence to the maximum likelihood. The upper panel compares the one-parameter marginalized Ω_m constraint for full numerical marginalization (black solid line) with analytic marginalization using equation (34) (red dot–dashed line), the difference between these lines, even in this non-Gaussian case, is small. The dashed horizontal lines show the one-parameter 1σ, 2σ and 3σ limits (assuming a Gaussian likelihood).

Since the galaxy bias parameter is poorly known, it is useful to marginalize over the amplitude when estimating Ω_m. The upper plot in Fig. 1 shows the projected 1D marginalized likelihood for Ω_m, for both numerical marginalization over the amplitude (black line) and using the analytic marginalization result given by equation (34) (red line). The analytic result accurately reproduces the full numerical result for the 1σ, 2σ and 3σ errors, even though there is some non-Gaussianity in the Ω_m–A plane.

4.2 Projection of parameter space

Another application for analytic marginalization is in the projection of parameter space. Usually the maximum likelihood parameter values are quoted along with the marginalized errors and marginalized parameter covariances. Sometimes the mean of a parameter, marginalized over all other parameters, is also quoted (e.g. Spergel et al. 2003), and the 2D projected parameter space plotted to illustrate non-Gaussianity. We can again use analytic marginalization to do this for us.

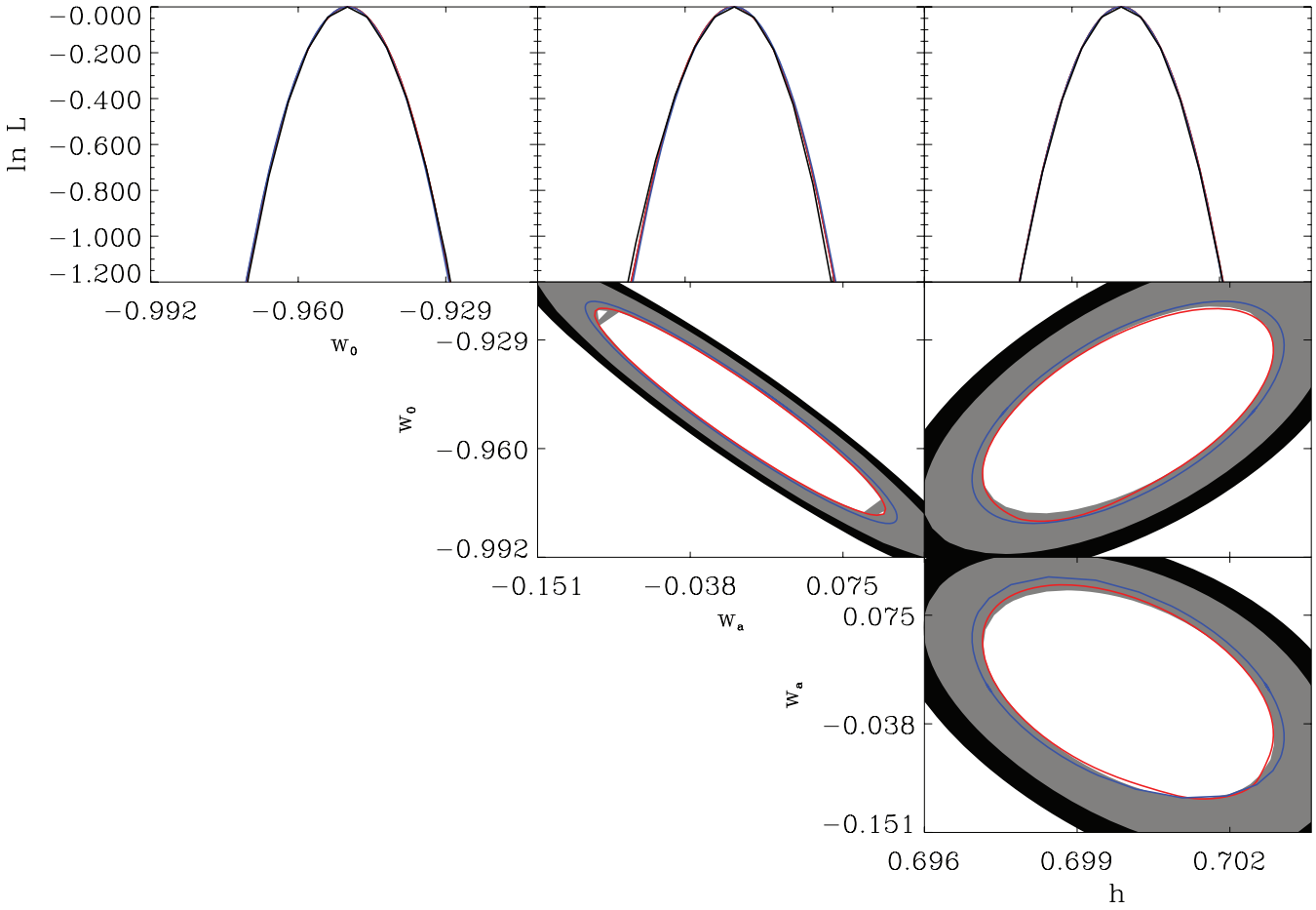

Figure 2. Projected cosmological three-parameter space for a Euclid-type (20 000 deg^2, median redshift of $z = 0.8$) gravitational lensing survey. Grey contours are 1σ 2σ and 3σ levels using analytic marginalization over the extra parameters, solid blue lined ellipses are the 1σ contours using the Fisher matrix approximation to the projected likelihood surface, solid red lined ellipses are the 1σ fully marginalized constraints. The upper panels show the 1D marginalized likelihoods for the analytic marginalization (black lines), the Fisher approximation (blue lines) and for a full numerical marginalization (red lines) – in the upper panel all the lines are effectively coincident, which highlights the accuracy of the analytic marginalization approach in such cases.

4.2.1 Dark energy parameters from 3D cosmic shear

In Fig. 2, we show the predicted projected likelihood space estimated on a grid for a set of three cosmological parameters (w_0, w_a, h), where $w(a) = w_0 + (1 - a)w_a$ is the dark energy equation of state, $p = w(a)\,\rho$ and $h = H_0/100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is the reduced Hubble parameter. The fiducial maximum-likelihood values are $w_0 = -0.95, w_a = 0$ and $h = 0.7$, and we have assumed a 3D tomographic cosmic shear analysis with the proposed Euclid satellite mission (Refregier et al. 2010), covering 20 000 deg^2 with median redshift $z = 0.8$. The upper row in Fig. 2 compare the analytically marginalized 1D parameter distribution with numerical marginalization over the remaining 2D likelihood surface and the Fisher matrix prediction. We see that analytic marginalization is indistinguishable from numerical marginalization. The lower panels show the projected 2D likelihood surface for analytic marginalization (solid white/grey/black 1σ, 2σ, 3σ regions) along with the two-parameter 1σ (68.3 per cent) likelihood contours estimated from the Fisher matrix approximation (blue ellipse), and a contour for the numerical marginalization (red ellipse). It can be seen in all panels that the analytic marginalized likelihood surface is in excellent agreement with the numerical marginalization, reproducing even small departures from the Fisher matrix approximation. While results will clearly depend on which parameters are in the likelihood analysis, this does suggest that for large numbers of parameters, the marginalization will tend towards a Gaussian distribution, since any departures from Gaussianity will be averaged out.

In Fig. 3 we extend the comparison to an eight-parameter cosmological model. In this example, the qualitative differences between the analytic marginalization results are clear. In some 2D parameter spaces for example (Ω_b, h) there is significant non-Gaussianity, however in others such as (w_0, w_a) the 2D parameter space is very Gaussian. In such circumstances, analytic marginalization could be used to marginalize over Gaussian parameter combinations and a numerical marginalization used to capture any non-Gaussian behaviour.

4.3 Semi-analytic marginalization

Non-Gaussianity may be significant for some parameters and so we propose an algorithm for *semi-analytic marginalization*. In this section, we propose two algorithms for this: *top-down* and *bottom-up*. In the bottom-up approach we first find the $(N_\mathrm{p} + M)$-parameter maximum-likelihood peak by a quasi-Newton solution,

$$\delta\Phi_\nu = -\frac{1}{2} F^{-1}_{\mu\nu} \mathcal{L}_\mu \tag{42}$$

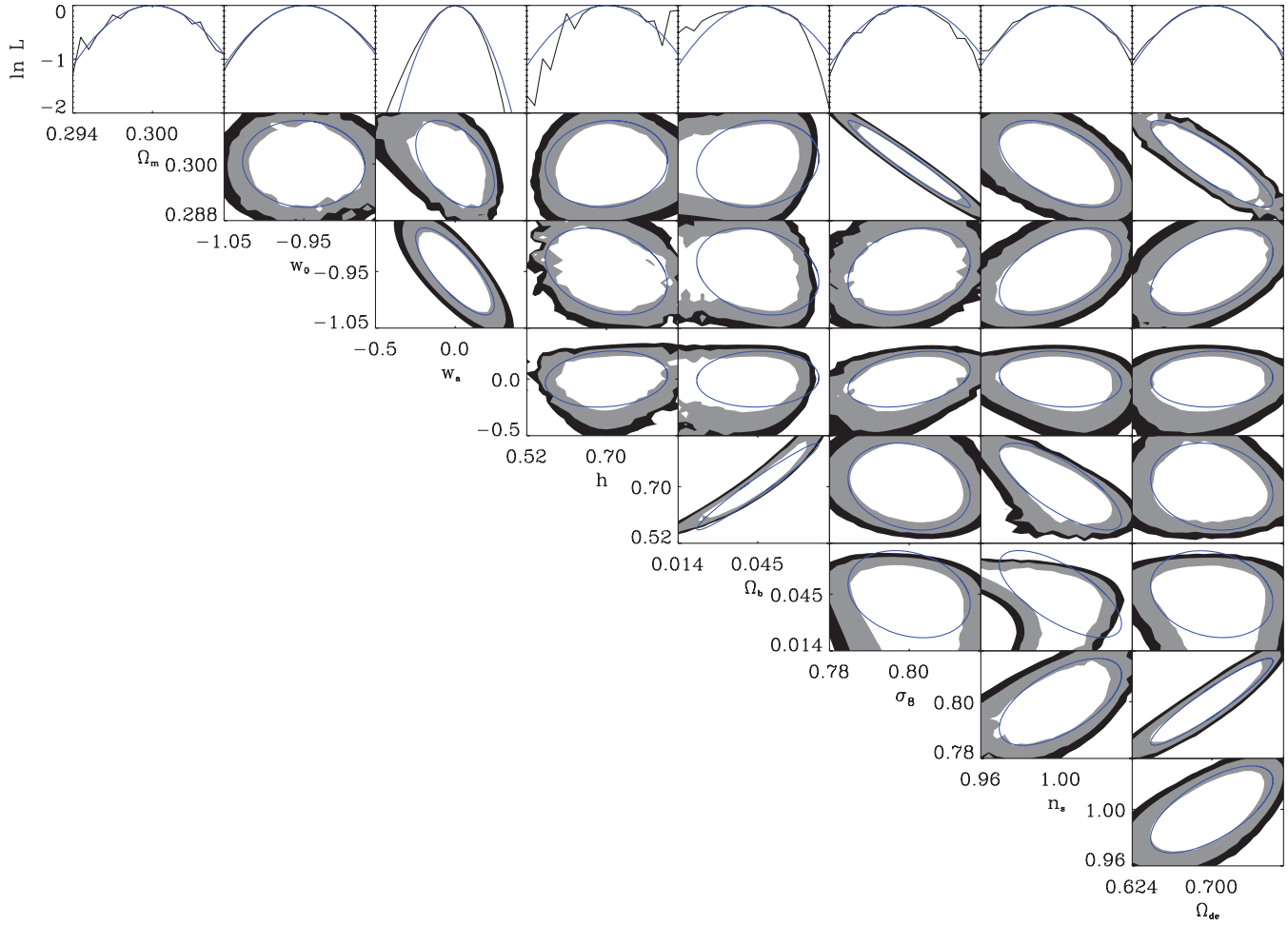

Figure 3. Projected cosmological eight-parameter space for a Euclid-type (20 000 deg^2, median redshift of $z = 0.8$) gravitational lensing survey. The upper panels show the 1D parameter constraints using analytic marginalization (black) and the Fisher matrix approximation (blue, dark grey). The other panels show the 2D parameter constraints. Grey contours are 1σ, 2σ and 3σ levels using analytic marginalization over the extra parameters, (blue) solid-lined ellipses are the 1σ contours using the Fisher matrix approximation to the projected likelihood surface.

and use MCMC to plot out the 1D and 2D parameter likelihood distributions, analytically marginalized over all other parameters. Any non-Gaussian parameters can be removed from the analytic marginalization and numerically marginalized over with MCMC. If new, non-Gaussian parameters appear we can numerically marginalize over them until stability is reached. A potential disadvantage of this approach is that non-Gaussian features may be obscured by projection which may bias confidence regions. We will investigate this algorithm elsewhere. This process may also end up running MCMC on all parameters – but in many cases some, if not many, of the parameters will be close to Gaussian-distributed in parameter space with just a few non-Gaussian parameters needing numerical marginalization.

An alternative, and more robust, approach is top-down where we would run an initial short-chain MCMC analysis and identify the Gaussian directions. We would then run subsequent semi-analytic marginalization, analytically marginalizing over the Gaussian parameters, and MCMC chains over the raining parameters. While this version clearly requires an initial, full MCMC analysis to be run, the speed-up will occur if we have to run the parameter estimation analysis many times as is common in data analysis or is run on a series of simulated data. However, if the number of parameters is too large, a traditional MCMC approach may not be feasible, in which case a bottom-up approach will be required.

Once the Gaussian directions are identified, the time spent mapping parameter space can be decreased significantly. We assume the time to run a full MCMC analysis in a N_p-parameter space is

$$T_{MC} = \Delta t_{MC} N_p \ln N_p, \tag{43}$$

where Δt_{MC} is the time to run one point in the MCMC chain. If M of these parameters can be analytically marginalized over, a semi-analytic marginalization scheme will take

$$T_{SAM} = \Delta t_{MC}(N_p - M)\ln(N_p - M) + \Delta t_F M, \tag{44}$$

where $\Delta t_F \ll \Delta t_{MCMC}$ is the time taken to estimate the Fisher matrix. Clearly if all parameters are well approximated by a multivariate Gaussian, the main effort is in finding the peak of the likelihood, since we already know the Fisher matrix. For example in our eight-parameter cosmological model (Fig. 3), only the baryon density, Ω_b, and the scalar spectral index, n_s, show significant deviations from Gaussianity. This implies we can reduce the computation time by a factor of 12. If we have a model with an additional 200 nuisance parameters, all of which can all be marginalized over, this is a reduction of around 67. Even if MCMC has be to extensively used to map out the parameter space, analytic marginalization can also

be used to map the MCMC proposal distributions more accurately than a Fisher matrix approximation.

5 MODEL SELECTION AND THE BAYESIAN EVIDENCE

5.1 The Bayesian evidence

Having explored analytic methods for maximizing and marginalizing in a likelihood analysis, we now turn to the problem of model selection. For model selection, we need to find the probability of the most likely model given the data, $p(\mathcal{M}|\boldsymbol{D})$. From Bayes' theorem, we find (see, e.g., Trotta 2008; Liddle 2009)

$$p(\mathcal{M}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\mathcal{M})p(\mathcal{M})}{p(\boldsymbol{D})}, \tag{45}$$

where the probability $p(\boldsymbol{D}|\mathcal{M})$ can be identified as the evidence from the likelihood analysis (equation 1). The probability $p(\mathcal{M})$ is the prior probability of the model in the absence of the data, for example from a previous experiment. The evidence, the probability of getting the data given the model for the system, is found by marginalizing over all cosmological parameters in the model,

$$E(\boldsymbol{D}|\mathcal{M}) = p(\boldsymbol{D}|\mathcal{M}) = \int \mathrm{d}^{N_\mathrm{p}}\theta\, L(\boldsymbol{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}). \tag{46}$$

This can be estimated numerically using thermodynamic integration (Slosar et al. 2003; Beltran et al. 2005), a variant of MCMC, or by nested sampling (Skilling 2004; applied to cosmology by Bassett, Corasaniti & Kunz 2004; Mukherjee et al. 2006) or VEGAS, a multi-dimensional integrator developed in particle physics (Lepage 1978) and applied in cosmology by Serra, Heavens & Melchiorri (2007). Alternative, approximate methods are the Savage–Dickey ratio for nested models (Trotta 2007), and the Bayesian information criterion (Schwartz 1987). When combining independent data set, parameter estimation only requires the addition of the log-likelihoods, but the Bayesian evidence must be re-evaluated by marginalization over the product of the posteriori distributions. For a large parameter space, the estimation of the evidence can be highly CPU-intensive, and so analytic methods are desirable.

5.1.1 The Laplace approximation

There is already a well-known analytic marginalization method which uses the saddle point, or Laplace, approximation (see e.g. MacKay 2003; Trotta 2008), where the likelihood is expanded around the peak in parameter space:

$$\mathcal{L}_\mathrm{Laplace} = \mathcal{L}_\mathrm{max} + \frac{1}{2}\Delta\theta_i\Delta\theta_j\mathcal{L}_{ij}, \tag{47}$$

where $\mathcal{L}_\mathrm{max}$ is evaluated at the maximum of the likelihood function in the full parameter space, and $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}_\mathrm{max}$. With a flat prior, $p(\theta|\mathcal{M}) = 1/V_\theta$, where V_θ is the prior volume of parameter space, we can carry out the Gaussian integration to find

$$\mathcal{L}_\mathrm{Laplace} = \mathcal{L}_\mathrm{max} + 2\ln(V_\theta\sqrt{\det F_{ij}}). \tag{48}$$

The last term is again the *Occam factor*, the ratio of the prior (non-zero) volume of parameter space to the effective posterior volume measured by the parameter covariance matrix, $\langle\Delta\theta_i\Delta\theta_j\rangle = F_{ij}^{-1}$.

A severe limitation of the Laplace approximation is that the value of $\mathcal{L}_\mathrm{max}$ is evaluated at the maximum likelihood point in parameter space,

$$\mathcal{L}_\mathrm{max} = \mathcal{L}(\theta_\mathrm{max}|\boldsymbol{D}, \mathcal{M}), \tag{49}$$

which depends on the data. Hence to evaluate it, we must first find the maximum likelihood for each model. To circumvent this, embedded or nested models have been considered, where the relative evidence between the evidence in one parameter space can be compared with that of a lower dimensional parameter space (see e.g. Heavens, Kitching & Verde 2007).

5.1.2 Analytic evidence

However, with analytic marginalization we now have a way to estimate the maximum of the likelihood for an arbitrary data set and fixed fiducial parameter values (Section 2.2). Expanding the cosmological parameter space to second order and marginalizing, and this time keeping all terms, we find

$$\mathcal{E} = \mathcal{L}_0 - \frac{1}{2}\mathcal{L}_i\mathcal{L}_{ij}^{-1}\mathcal{L}_j + \mathrm{Tr}\,\ln V_\theta^{2/N_\mathrm{p}}\mathcal{L}_{ij} - N_\mathrm{p}\ln 4\pi, \tag{50}$$

where $\mathcal{E} \equiv -2\ln E$ is the log-evidence. This expression is again independent of the fiducial model used, as we should expect after marginalization.

5.1.3 Gaussian likelihoods

If the likelihood for the data is Gaussian and the parameters appear in the mean, the evidence is

$$\mathcal{E}(\boldsymbol{D}|\mathcal{M}) = \Delta\boldsymbol{D}\left(\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{\mu}_i^t F_{ij}^{-1}\boldsymbol{\mu}_j\boldsymbol{C}^{-1}\right)\Delta\boldsymbol{D}^t + \mathrm{Tr}\,\ln\boldsymbol{C} + 2\ln(V_\theta\sqrt{\det F_{ij}}) - N_\mathrm{p}\ln 2\pi. \tag{51}$$

The evidence is the probability based on the outcome of given experiment. However we can also forecast the evidence of future experiments and ask what is the expected evidence, and even what is the variance on a prediction of the evidence. Just like the frequentist χ^2-statistic, this will give us an expectation of the mean and range of values of evidence we should expect from an experiment, given the uncertainty in the data.

The expectation value of the Gaussian log-evidence is

$$\langle\mathcal{E}\rangle = \nu + \mathrm{Tr}\,\ln\boldsymbol{C} + 2\ln(V_\theta\sqrt{\det F_{ij}}) - N_\mathrm{p}\ln 2\pi, \tag{52}$$

where $\nu = N_\mathrm{D} - N_\mathrm{p}$ is the number of degrees of freedom, N_D is the number of data points and N_p is the number free parameters. This is then just the χ^2 number of degrees of freedom, plus the normalization factor and the Occam factor. If we were to ignore these terms, we see the Gaussian log-evidence, $\mathcal{E}$, has the same expectation value as the χ^2-statistic. If we further estimate the variance of the log-evidence, we find

$$\langle\Delta\mathcal{E}^2\rangle = 2\nu \tag{53}$$

is just twice the number of degrees of freedom, as we might expect for a Gaussian distribution. This highlights the connection between the evidence and the χ^2-statistic, and shows that, although they are asking different questions of the data, they have a similar 'sensitivity'.

5.1.4 Evidence for an arbitrary model

In addition to calculating the evidence for the data, given the maximum likelihood model also from the data, we can also ask what is the probability that the measured data are drawn from an arbitrary model, given an assumed set of 'true' parameter values, $p(\boldsymbol{D}|\mathcal{M}_t)$,

and scatter in the possible data. We can calculate this from

$$\mathcal{E}(\boldsymbol{D}|\mathcal{M}_t) = \Delta \boldsymbol{D}\boldsymbol{C}^{-1}\Delta \boldsymbol{D}^t + \mathrm{Tr}\,\ln \boldsymbol{C} \tag{54}$$

$$+ 2\ln(V_\theta \sqrt{\det F_{ij}}) - N_\mathrm{p}\ln 2\pi, \tag{55}$$

where the likelihood peaks at the 'true' values, not the values which best fit the data. As an example, if the maximum likelihood given the data peaks at a non-ΛCDM (non-standard model), equation (51) will yield the evidence for that model. But instead if we assume that ΛCDM parameters is the 'true' model; equation (54) will tells us the probability that the data are drawn from this model. If this is very low, it is unlikely the data are drawn from this model.

5.1.5 The Occam factor

The final term in the evidence, the Occam factor, is often problematic as it depends on the assumed prior volume of the parameter space, which is not well defined. While we can hope that for good data the other terms in the evidence dominate over the Occam factor, for poor data, this may not be the case. One approach is to assume that the prior is set using the Fisher matrix. We can let $V_\theta = a^{N_\mathrm{p}}/\sqrt{\det F_{ij}}$, where the constant of proportionality of order $a = 10$ and N_p is the number of parameters. This factor becomes simply $2N_\mathrm{p}\ln a$, and so this term still gives more weight to models with fewer parameters. The parameter a becomes an adjustable parameter, depending on how much weight one wants to give to the Occam factor. A value of $a = 10$ would seem to be fairly conservative. Clearly, this scheme can be extended for parameter which are highly unconstrained.

We also note that our expression for the evidence will disfavour models which have arbitrary unconstrained parameters. A common concern in evidence calculations is that an extra parameter entirely unconstrained by the data could be added that would result in the disfavourment of the model only via the Occam factor. We find that in such an unconstrained model, the χ^2 term becomes infinity because the Fisher matrix element for these parameters is zero and hence the probability of such models is zero.

5.2 Model selection

5.2.1 Model selection: Bayes' factor

A common approach to model selection is the use of the Bayes' factor (Kass & Raftery 1995), the ratio of pairs of models or its logarithm,

$$\mathcal{B}_{AB} = -2\ln B_{AB} = \mathcal{E}(\boldsymbol{D}|\mathcal{M}_A) - \mathcal{E}(\boldsymbol{D}|\mathcal{M}_B). \tag{56}$$

This has the advantage that we do not need to consider the normalization factor, $p(\boldsymbol{D})$, in Bayes' equation (45). Jeffreys (1961) has proposed a qualitative scale based on these ratios.

5.2.2 Model selection: Model-space

An alternative is to rank-order models by their evidence, with a uniform prior, $p(\mathcal{M}) = 1/N_\mathrm{M}$, where N_M is the number of models. Even though we do not expect to have a complete set of all possible models, we can still normalize the set we have to estimate the posterior probability for each model, $\mathcal{M}_A$:

$$p(\mathcal{M}_A|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\mathcal{M}_A)p(\mathcal{M}_A)}{\sum_B^{N_\mathrm{M}} p(\boldsymbol{D}|\mathcal{M}_B)p(\mathcal{M}_B)}, \tag{57}$$

where we consider independent models to form a countable set. By this definition, uncountable sets of models contain models that can be distinguished by a continuous parameter, which is then just a model with a variable parameter, i.e. we class a model as the set of parameters, not a set of parameter values.

Even though the models may be incomplete, $p(\mathcal{M}_A|\boldsymbol{D})$ is an upper limit on the true probability for each model with this data set. Adding any new model will only reduce the probability. Since the prior is uniform, we expect a new model to appear at random in the distribution. This scheme not only assesses 'goodness-of-fit' to the data, but also the competitiveness of models. If one model does well compared to other proposed models, we rightly attach more belief to it. However, it does not prevent a new model appearing with a higher evidence which would become the best model. In this scheme, one would not necessarily truncate or throw away models, since they contribute to the normalization of the probabilities – although if the contribution is negligible it would seem sensible to drop outliers so the model-space is of a manageable size.

5.2.3 Model significance

Even though the scheme outlined above puts an upper limit on the absolute model probability, it will still return the following result: that if we only have one model, Bayes' theorem tells us we must assign it a 100 per cent probability (since it is the only viable model). Instead we could judge a model in relation to the prior we assign it. To do this, we define a significance factor,

$$\mathcal{S} = \frac{p(\mathcal{M}|\boldsymbol{D})}{p(\mathcal{M})} = \frac{p(\boldsymbol{D}|\mathcal{M})}{p(\boldsymbol{D})}, \tag{58}$$

where, by definition, $\mathcal{S} \geq 1$, since we cannot lose information by adding data. The evidence for any model is only *significant* if the ratio, $\mathcal{S}$, of the evidence to the prior for the model $\mathcal{M}$ is much larger than unity. For example, if we consider again the situation when we only have one model the prior probability is $p(\mathcal{M}) = 1$, so that $\mathcal{S} = 1$, and we have not learned anything about the absolute validity of the model.

We can now estimate the number of models needed for any model to be convincing in an absolute sense. For two models the uniform prior for each model is $p(\mathcal{M}_A) = 1/2$, so the maximum significance is 2. While the Bayes' factor between the two models could 'decisively' favour one model over the other (odds of $\gtrsim$1:100 on Jeffreys scale), one could only be at most 'inconclusive' (odds of 1:2) that the model is correct. For absolute confidence, we need at least three models for comparison.[2] This argument can be used to retrospectively understand the history of model selection. For example, when given the choice of a steady state model over the big bang the later was clearly favoured due to a large Bayes' factor. However the absolute confidence in the big bang could not be high since there were no alternative theories. Indeed once inflationary cosmologies appeared this new theory became preferable.

If a new model is added to the model-space, the significance, $\mathcal{S}_A$ scales as

$$\mathcal{S}'_A = \frac{\mathcal{S}_A(N_\mathrm{M}+1)}{N_\mathrm{M} + \mathcal{S}_A[p(\boldsymbol{D}|\mathcal{M}_\mathrm{new})/p(\boldsymbol{D}|\mathcal{M}_A)]}. \tag{59}$$

If the new model has lower probability the significance scales as $\mathcal{S}'_A = \mathcal{S}_A(N_\mathrm{M}+1)/N_\mathrm{M}$, while if it has much higher probability it scales as $\mathcal{S}'_A = (N_\mathrm{M}+1)p(\boldsymbol{D}|\mathcal{M}_A)/p(\boldsymbol{D}|\mathcal{M}_\mathrm{new})$.

[2] Note the prior on the model is important here. A flat prior of $1/N_\mathrm{M}$ is only appropriate for equally credible models. Including a vast array of non-credible models can be countered by giving these a low-prior weighting.

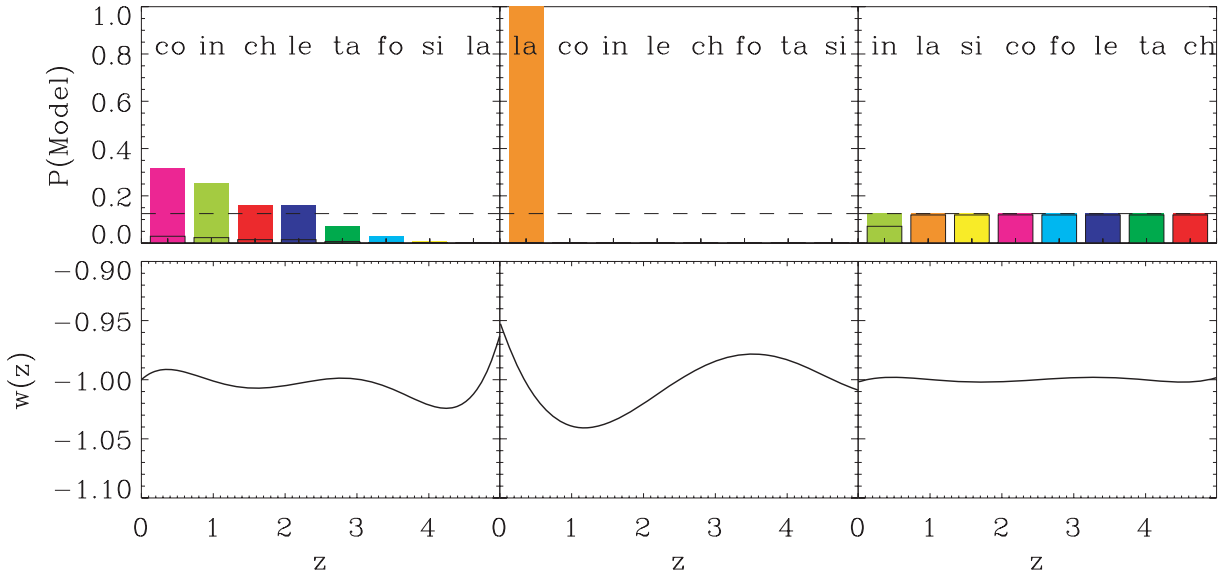

Figure 4. A simple example of non-nested evidence analysis. The bottom row shows three $w(z)$ realizations, the top row shows the corresponding rank-ordered, non-nested evidence for each model on the left (using a Euclid weak lensing tomography experiment). The models are fo=Fourier (turquoise), ch=Chebyshev (red), la=Laguerre (orange), Le=Legendre (blue), in=Interpolation (dark green), ta=Taylor (light green), co=Cosine (purple) and si=Sine (yellow) (see Kitching & Amara 2009, for details). These represent the three possible classes of expected model-space, a broad variance but with a favoured model; a highly favoured model; or a broad set of equally favoured models. In solid outlined bars we show the evidence that the data are drawn from a ΛCDM cosmology instead of the best-fitting values to the data. The dashed line show the flat model prior, $p(\mathcal{M}) = 1/N_{\rm M}$.

5.2.4 Dark energy model-space

In Fig. 4, we show an example of how the evidence can be used in practice for the predicted evidence for a Euclid (Refregier et al. 2010) weak lensing tomography experiment to measure dark energy. In this example, we have assumed a dark energy equation of state, $w(z)$, as a function of redshift, z, which we use to construct mock lensing data. We fit this data using models that assume a cosmology with different $w(z)$ models. We have chosen some non-nested basis set expansions for our $w(z)$ models; these have a maximum order of 2 (these phenomenological models are described in Kitching & Amara 2009). For each $w(z)$ realization, we rank-order the evidence for each model. In the first example, the cosine model has the highest probability with 0.4 and the distribution in model-space is Gaussian-like. In the second example, the Chebyshev model fits the data very well, creating a spike in model-space. In the third example, there is no model that favours the data over any other. These three examples represent the three broad classes of behaviour we can expect for real data, where we hope for example 2 with a spike in model-space. The variance in model-space is also an interesting quantity, reflecting both the distinguishability of the models and the quality of the data for model selection.

6 DISCUSSION

We have presented new, analytic methods for cosmological likelihood analysis to solve the 'many parameters' problem in cosmology. Our approach maximizes the likelihood with a pseudo-Newton method, analytically marginalizes over nuisance parameters in an arbitrary likelihood function, and analytically marginalizes over cosmological parameters to project out one and two dimensions of parameter space to estimate marginalized errors and covariance matrices. Parameters may have either flat or Gaussian priors. Marginalizing over all parameters, we derive an analytic expression for the Bayesian evidence to select between competing cosmological models. The marginalized likelihood does not degrade information about the remaining parameters, and the marginalized parameter information is preserved in the Fisher information matrix. The marginalized likelihood is also independent of the fiducial model when the underlying likelihood is exactly Gaussian.

We have applied our results to multivariate Gaussian likelihoods for the data, where the marginalized parameters appearing in either the mean of the data or its covariance matrix. An exact result for a normalization nuisance parameter is found and applied to the problem of estimation the matter density parameter, $\Omega_{\rm m}$, from galaxy power spectra, where the normalization, which depends on the galaxy bias parameter, b, is marginalized out. The analytic marginalization is found to be very close to numerical marginalization. Analytic marginalization can also be used to project parameter space on to lower dimensions to allow a simple visualization of the full likelihood function.

We describe a semi-analytic marginalization method which could be carried out by identifying Gaussian and non-Gaussian parameters, in top-down or bottom-up scenarios, and treating them analytically and numerically, respectively, in semi-analytic marginalization. An example is presented of a three-parameter dark energy model with $(w_0, w_{\rm a}, h)$, and again the 1D analytically marginalized distribution is in very good agreement with the numerical one. We extend this to an eight-parameter model, where we highlight non-Gaussianity in the 2D projected distribution which is missed by the Fisher matrix approximation.

Finally, we have also applied our analytic marginalization method to find a closed expression for the Bayesian evidence and shown its relation to the Laplace approximation. We discuss the case of multivariate Gaussian-distributed data sets. We consider the Bayes' factor, the ratio of the evidence of two models, and discuss the properties of the full model-space posteriori distribution, $p(\mathcal{M})$. We also introduce the significance of the model, the degree by which the model evidence changes with respect to the uniform prior. Finally, we have illustrated our model selection scheme on a set of

non-nested dark energy models. Our method has applications in cosmological parameter estimation and model selection, and many wider applications in the statistical analysis of data.

ACKNOWLEDGMENTS

We thank Andrew Liddle, John Peacock, Alan Heavens, Fergus Simpson, Adam Amara and Benjamin Joachimi for much useful discussion. We also thank the DUEL network (MRTN-CT-2006-036133) for supporting part of this work. TDK is supported by STFC rolling grant number RA0888.

REFERENCES

Bassett B. A., Corasaniti P. S., Kunz M., 2004, ApJ, 617, L1
Beltran M., Garcia-Bellido J., Lesgourgues J., Liddle A. R., Slosar A., 2005, Phys. Rev. D, 71, 063532
Bretthorst G., 1988, Lecture Notes Ser. Vol. 48, Bayesian Spectrum Analysis and Parameter Estimation. Springer-Verlag, New York
Bridle S. L., Crittenden R., Melchiorri A., Hobson M. P., Kneissl R., Lasenby A. N., 2002, MNRAS, 335, 1193
Eisenstein D., Hu W., 1997, ApJ, 511, 5
Gamerman D., 1997, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall, New York
Gull S. F., 1989, in Skilling J., ed., Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, p. 511
Heavens A. F., Taylor A. N., 1995, MNRAS, 275, 483
Heavens A. F., Kitching T. D., Taylor A. N., 2006, MNRAS, 373, 105
Heavens A. F., Kitching T. D., Verde L., 2007, MNRAS, 380, 1029
Jeffreys H., 1961, Theory of Probability, 3rd edn., Oxford Univ. Press, Oxford
Kaiser N., 1988, MNRAS, 231, 149
Kass R. E., Raftery A. E., 1995, J. Am. Statistical Association, 90, 773
Kitching T., Amara A., 2009, 398, 2134
Kosowsky A., Milosavljevic M., Jimenez R., 2002, Phys. Rev. D, 66, 063007
Lepage G. P., 1978, J. Comput. Phys., 27, 192
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
Liddle A. R., 2009, Annu. Rev. Nuclear Part. Sci., 59, 95
MacKay D. J. C., 2003, Information theory, inference and learning algorithms. Cambridge Univ. Press, Cambridge
Mukherjee P., Parkinson D., Corasaniti P. S., Liddle A. R., Kunz M., 2006, MNRAS, 369, 1725
Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T., 1989, Numerical Recipes: The Art of Scientific Computing. Cambridge Univ. Press, Cambridge
Refregier A., Amara A., Kitching T. D., Rassat A., Scaramella A., Weller J., 2010, preprint (arXiv:1001.0061)
Schwartz G., 1987, Ann. Statist., 5, 461
Serra P., Heavens A., Melchiorri A., 2007, MNRAS, 379, 169
Skilling J., 2004, in Fischer R., von Toussaint U., Preuss R., eds, AIP Conf. Proc. Vol. 735, Baysian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 395
Slosar A. et al., 2003, MNRAS, 341, L29
Smith R. et al., 2003, MNRAS, 341, 1311
Spergel D. et al., 2003, ApJS, 148, 175
Taylor A. N., Watts P., 2001, MNRAS, 328, 1027
Taylor A. N., Kitching T., Bacon D., Heavens A., 2007, MNRAS. 374, 1377
Tegmark M., 1997, Phys. Rev. D, 55, 5895
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
Tegmark M. et al., 2004, Phys. Rev. D, 69, 103501
Trotta R., 2007, MNRAS, 378, 72
Trotta R., 2008, Contemporary Phys., 49, 71
Verde L. et al., 2003, ApJS, 148, 195
Woodbury M. A., 1950, Inverting modified matrices, Memorandum Rept. 42, 4. Statistical Research Group, Princeton Univ. Press, Princeton, NJ
Zhang F., 2005, The Schur Complement and its Applications. Springer, New York

APPENDIX A: GAUSSIAN INTEGRATION

In this appendix we derive equation (7). Expanding the log-likelihood to second order we find

$$\mathcal{L} = \mathcal{L}_0 + \delta\psi_\alpha \mathcal{L}_\alpha + \frac{1}{2}\delta\psi_\alpha \delta\psi_\beta \mathcal{L}_{\alpha\beta}. \tag{A1}$$

By completing the square, this can be rewritten as

$$\mathcal{L} = \mathcal{L}_0 + \frac{1}{2}\mathcal{L}_{\alpha\beta}\left(\mathcal{L}_\gamma \mathcal{L}^{-1}_{\gamma\alpha} + \delta\psi_\alpha\right)\left(\mathcal{L}_\delta \mathcal{L}^{-1}_{\delta\beta} + \delta\psi_\beta\right) - \frac{1}{2}\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta. \tag{A2}$$

Now writing the likelihood explicitly, we find

$$L = \mathrm{e}^{-\frac{1}{2}\mathcal{L}_0 + \frac{1}{4}\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta - \frac{1}{4}\mathcal{L}_{\alpha\beta}(\mathcal{L}_\gamma \mathcal{L}^{-1}_{\gamma\alpha} + \delta\psi_\alpha)(\mathcal{L}_\delta \mathcal{L}^{-1}_{\delta\beta} + \delta\psi_\beta)}. \tag{A3}$$

Integrating over $\delta\boldsymbol{\psi}$, and using the multivariate Gaussian formula

$$\int \mathrm{d}^n x\, \mathrm{e}^{-\frac{1}{2}x_i C^{-1}_{ij} x_j} = (2\pi)^{n/2}\sqrt{\det \boldsymbol{C}}, \tag{A4}$$

we find

$$L = \mathrm{e}^{-\frac{1}{2}\mathcal{L}_0 + \frac{1}{4}\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta}(2\pi)^{N/2}\sqrt{\det 2\mathcal{L}^{-1}_{\alpha\beta}}. \tag{A5}$$

Taking the log again we find

$$\mathcal{L} = \mathcal{L}_0 - \frac{1}{2}\mathcal{L}_\alpha \mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta + \ln\det\frac{1}{2}\mathcal{L}_{\alpha\beta} - N\ln 2\pi. \tag{A6}$$

Using the identity $\ln\det M = \mathrm{Tr}\,\ln M$ yields equation (7).

APPENDIX B: GENERATING FUNCTION

The generating function of a distribution is

$$\Phi(\boldsymbol{J}) = \langle \mathrm{e}^{i\boldsymbol{J}.\delta\boldsymbol{\psi}}\rangle = \int \mathrm{d}^M\psi\, \mathrm{e}^{-\mathcal{L}/2}\, \mathrm{e}^{i\boldsymbol{J}.\delta\boldsymbol{\psi}} \tag{B1}$$

which leads to the generating function of the likelihood:

$$-2\ln\Phi(J) = \mathcal{L}_0 - \frac{1}{2}(\mathcal{L}_\alpha - 2iJ_\alpha)\mathcal{L}^{-1}_{\alpha\beta}(\mathcal{L}_\beta - 2iJ_\beta) + \mathrm{Tr}\,\ln\frac{1}{2}\mathcal{L}_{\alpha\beta}. \tag{B2}$$

Taking the first derivative with respect to iJ_α, we find the mean is

$$\langle\delta\psi_\alpha\rangle = \left.\frac{\partial\ln\Phi}{\partial(iJ_\alpha)}\right|_{J=0} = -\mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_\beta(\boldsymbol{\theta}). \tag{B3}$$

For a Gaussian the mean is also at the peak, so this is an offset between a fixed point, $\boldsymbol{\psi}_0$, where the likelihood is evaluated and the peak. The second derivative yields the covariance matrix

$$\langle\delta\psi_\alpha\delta\psi_\beta\rangle = \left.\frac{\partial^2\ln\Phi}{\partial(iJ_\alpha)\partial(iJ_\beta)}\right|_{J=0} = 2\mathcal{L}^{-1}_{\alpha\beta}. \tag{B4}$$

Taking the ensemble average of the data, we see

$$\langle\delta\psi_\alpha\delta\psi_\beta\rangle = F^{-1}_{\alpha\beta} \tag{B5}$$

as expected. Expanding θ around its maximum-likelihood value, we find

$$\langle\delta\psi_\alpha\rangle = -\mathcal{L}^{-1}_{\alpha\beta}\mathcal{L}_{\beta i}\Delta\theta_i. \tag{B6}$$

Finally, inverting this we find the bias in cosmological parameters, $\delta\boldsymbol{\theta}$, due to an offset in the nuisance parameter is given by

$$\delta\theta_i = -\mathcal{L}^{-1}_{i\alpha}\mathcal{L}_{\alpha\beta}\delta\psi_\beta. \tag{B7}$$

in agreement with the result of Taylor et al. (2007).

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