Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **510,** 3207–3221 (2022) https://doi.org/10.1093/mnras/stab3540
Advance Access publication 2021 December 4

Matching Bayesian and frequentist coverage probabilities when using an approximate data covariance matrix

Will J. Percival,[1,2,3]★ Oliver Friedrich,[4,5] Elena Sellentin[6,7] and Alan Heavens[8]

[1] *Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[2] *Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[3] *Perimeter Institute for Theoretical Physics, 31 Caroline St. North, Waterloo, ON N2L 2Y5, Canada*
[4] *Kavli Institute for Cosmology, University of Cambridge, Cambridge CB3 0HA, UK*
[5] *Churchill College, University of Cambridge, Cambridge CB3 0DS, UK*
[6] *Mathematical Institute, Leiden University, Snellius Gebouw, Niels Bohrweg 1, NL-2333 CA Leiden, the Netherlands*
[7] *Leiden Observatory, Leiden University, Oort Gebouw, Niels Bohrweg 2, NL-2333 CA Leiden, the Netherlands*
[8] *Imperial Centre for Inference and Cosmology (ICIC), Department of Physics, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*

Accepted 2021 November 30. Received 2021 November 24; in original form 2021 August 17

ABSTRACT

Observational astrophysics consists of making inferences about the Universe by comparing data and models. The credible intervals placed on model parameters are often as important as the maximum a posteriori probability values, as the intervals indicate concordance or discordance between models and with measurements from other data. Intermediate statistics (e.g. the power spectrum) are usually measured and inferences are made by fitting models to these rather than the raw data, assuming that the likelihood for these statistics has multivariate Gaussian form. The covariance matrix used to calculate the likelihood is often estimated from simulations, such that it is itself a random variable. This is a standard problem in Bayesian statistics, which requires a prior to be placed on the true model parameters and covariance matrix, influencing the joint posterior distribution. As an alternative to the commonly used independence Jeffreys prior, we introduce a prior that leads to a posterior that has approximately frequentist matching coverage. This is achieved by matching the covariance of the posterior to that of the distribution of true values of the parameters around the maximum likelihood values in repeated trials, under certain assumptions. Using this prior, credible intervals derived from a Bayesian analysis can be interpreted approximately as confidence intervals, containing the truth a certain proportion of the time for repeated trials. Linking frequentist and Bayesian approaches that have previously appeared in the astronomical literature, this offers a consistent and conservative approach for credible intervals quoted on model parameters for problems where the covariance matrix is itself an estimate.

Key words: methods: data analysis – methods: statistical – cosmology: observation.

1 INTRODUCTION

The problem of fitting a model to multivariate Normal (hereafter referred to as Gaussian) distributed data, where only an approximation to the true data covariance matrix is available, often arises in astrophysics. In a Bayesian sense, the problem can be considered as jointly fitting a model for the data and the covariance matrix, which is a standard one in statistics with a long history. For Gaussian-distributed data, the standard estimate of the covariance matrix is drawn from a Wishart distribution, such as when a covariance matrix is estimated using a limited number of simulations, or when a covariance matrix is constructed from Jackknife samples (e.g. Norberg et al. 2009; Friedrich et al. 2016). Examples of cosmological inferences made within this framework include the recent measurements from BOSS and eBOSS (Alam et al. 2017; eBOSS Collaboration 2020) as well as the galaxy clustering part of Heymans et al. (2021). For analyses of two-point statistics in line-of-sight projected data the covariance matrix is often modelled analytically instead of estimating it from simulations (see e.g. Krause & Eifler 2017; Heymans et al. 2021; DES Collaboration 2021, for recent examples). This is because the four-point functions constituting those covariances are accurately approximated in a Gaussian model, that is easy to evaluate (Friedrich et al. 2021; Joachimi et al. 2021). In contrast, analyses of non-standard summary statistics almost exclusively rely on estimated covariances because analytical covariance models are not easily obtained for them (e.g. Kacprzak et al. 2016; Brouwer et al. 2018; Gruen et al. 2018; Martinet et al. 2018; Halder et al. 2021).

There are two common ways to characterize our uncertainty about a model parameter when comparing data and model, which lie at the heart of the difference between Bayesian and frequentist approaches. One can perform a Bayesian analysis using the posterior to define credible intervals, within which a model parameter falls with a particular probability given the prior information and experimental data. One can also define a mechanism to produce

★ E-mail: will.percival@uwaterloo.ca

© 2021 The Author(s)
Published by Oxford University Press on behalf of Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

frequentist confidence regions, a set proportion of which contain the true parameters in repeated trials. For astrophysical problems we can consider the trials to be experiments performed in parallel universes that are independent and identically distributed realizations of the same data generating process (so the universal constants are considered the same). Confidence regions determined, for example, by the distribution of the difference between truth and the maximum likelihood solution, will not in general be the same as the credible regions, and it is self-evidently wrong to identify them for asymmetric distributions (see e.g. Loredo 2012). That they are not generally the same is evident since credible regions are clearly dependent on the prior, while maximum likelihood estimates are not. In other words: the fraction of times the credible intervals contain the *true* parameters for repeated analyses (the frequentist coverage probability) is not necessary equal to the posterior probability enclosed within these intervals. The difference has previously been used in astrophysics to search for unrecognized biases during data analysis (Sellentin & Starck 2019).

In this paper, we seek a prior that gives a frequentist matching posterior, so that we can define credible regions that have the property that, for a given parametrization, the x per cent credible regions contain the true parameter values in approximately x per cent of repeated trials. This means that we can interpret the mechanism used to define these regions (the Bayesian mechanism) as providing confidence regions with a frequentist coverage probability that matches the Bayesian probability associated with interpreting the same regions as credible regions. This match always holds in the asymptotic limit of infinite data (the Bernstein–von Mises theorem), which includes having a perfect covariance matrix estimate; here our prior ensures the distributions match at the level of equal parameter covariances, for Gaussian linear models and approximately for nonlinear models.

Note that, in general, frequentist matching priors are not a panacea, as they may not perform well in all circumstances, such as in making predictive distributions (Sun & Berger 2006), and they are not invariant to reparametrization. Note also that the differences between the different priors diminish, as expected, when the number of simulations is large and the posterior is dominated by data.

Before we introduce the problem further and the frequentist matching solution, we introduce the notation adopted: $\boldsymbol{x}_0$ are the compressed experimental data of dimension n_d (e.g. a power spectrum), while $\boldsymbol{x}_i$ is the simulated data with $1 \le i \le n_s$, assumed to be Gaussian distributed around the true model. From the n_s simulations, we construct an unbiased estimate of the covariance matrix S,

$$S = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T, \tag{1}$$

where $\bar{\boldsymbol{x}}$ is the mean of $\boldsymbol{x}_i$ over all simulations. The expectation value of $\boldsymbol{x}_0$ is $\boldsymbol{\mu}$, and Σ its (unknown) covariance. We only use the simulated data to calculate S, and so we consider the data to be $(\boldsymbol{x}_0, S)$. We will consider fitting a model with n_θ parameters $\boldsymbol{\theta}$, such that our model for the data is $\boldsymbol{\mu}(\boldsymbol{\theta})$, while the covariance matrix used to form the posterior remains of dimension n_d. Without loss of generality we shall assume that the expected values of μ and θ are zero, such that they can be ignored in our equations and we can, for example, write the covariance for estimates of $\hat{\boldsymbol{\theta}}$ as $\langle \hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T \rangle$.

Errors in the covariance matrix used to determine the likelihood have a number of effects on the inferences we make from the data, and particularly the credible intervals quoted in a Bayesian analysis. Hartlap, Simon & Schneider (2007) was the first to point out in the astronomical literature that, for S calculated using equation (1) and therefore drawn from a Wishart distribution with degrees of freedom $n_s - 1$ and scale matrix $\Sigma/(n_s - 1)$, S^{-1} is a biased estimator for the inverse covariance matrix Σ^{-1}, whereas $(hS)^{-1}$ is not, where

$$h = \frac{n_s - 1}{n_s - n_d - 2} \tag{2}$$

is commonly (by astronomers) called the Hartlap factor (although knowledge of this effect reaches at least as far back as Kaufman 1967). We discuss the application of the Hartlap factor further in Section 8.

Taking a frequentist stance, Dodelson & Schneider (2013) and Taylor & Joachimi (2014) showed that the nature of S has a strong effect on the confidence intervals derived based on the distribution of maximum a posteriori probability (MAP) model parameters (commonly called the best-fitting parameters). In fact, we will show later that for the priors and linear models that we consider, the maximum likelihood and MAP parameters are the same. So, we could have considered this distribution as the distribution of maximum likelihood solutions. However, as most analyses only work with the posterior, we simply refer to these as the MAP model parameters. Dodelson & Schneider (2013) provided a second-order calculation deriving the distribution of MAP model parameters recovered after repeated experiments, averaging over a set of estimated covariance matrices. This derivation is reviewed in Section 3.3. Percival et al. (2014) pointed out that the offset found by Dodelson & Schneider (2013) cannot be applied directly to change credible intervals as the average posterior from a set of repeated experiments itself depends on the distribution of S, and they provided a factor by which the credible intervals recovered assuming a Gaussian posterior could be adjusted to match the confidence intervals obtained from the distribution of MAP parameters recovered from mocks. This is discussed further in Section 7.

The Bayesian solution was introduced in the astronomical literature by Sellentin & Heavens (2016) based on the independence Jeffreys prior and marginalizing over the unknown covariance matrix. The resulting posterior has multivariate t-distribution form. The derivation follows from Bayes theorem, starting from the joint posterior

$$f(\boldsymbol{\mu}, \Sigma | \boldsymbol{x}_0, S) \propto f(\boldsymbol{x}_0, S | \boldsymbol{\mu}, \Sigma) f(\boldsymbol{\mu}, \Sigma), \tag{3}$$

where $f(\boldsymbol{\mu}, \Sigma)$ is the prior, and $f(\boldsymbol{x}_0, S | \boldsymbol{\mu}, \Sigma)$ the likelihood. Because of the independence of $\boldsymbol{x}_0$ and S, the likelihood can be written

$$f(\boldsymbol{x}_0, S | \boldsymbol{\mu}, \Sigma) = f(\boldsymbol{x}_0 | \boldsymbol{\mu}, \Sigma) f(S | \Sigma). \tag{4}$$

To make model inferences, we wish to know the distribution of the data-generating mechanism (or its parameters) given the data and S, which we can calculate by marginalizing over the true covariance:

$$f(\boldsymbol{\mu} | \boldsymbol{x}_0, S) = \int \mathrm{d}\Sigma \; f(\boldsymbol{\mu}, \Sigma | \boldsymbol{x}_0, S). \tag{5}$$

The key question in a Bayesian analysis performed under these conditions is the form for the prior proposed for μ and the covariance matrix. Sun & Berger (2006) listed a number of options for prior choices, including the Jeffreys prior,

$$f(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-\frac{n_d+2}{2}}, \tag{6}$$

and independence Jeffreys prior (adopted by Sellentin & Heavens 2016),

$$f(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-\frac{n_d+1}{2}}. \tag{7}$$

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

Giesser & Cornfield (1963) consider a range of priors

$$f(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-v}, \tag{8}$$

where v is an integer with $v \leq n_s$. Various other potential priors have also been introduced (e.g. Haar prior, right-Haar prior, left-Haar prior, Chang & Eaves 1990 reference prior) with more complicated forms. Each has advocates and interesting properties in various situations.

The prior that we introduce is a member of the class of frequentist matching priors (Lindley 1958; Welch & Peers 1963; Reid, Mukerjee & Fraser 2003), designed to match a posterior to frequentist expectations. A discussion of such priors is given in Ghosh (2011). Priors that match posterior predictive probabilities with the corresponding frequentist probabilities are attractive when constructing credible/confidence intervals. In general, matching priors can be constructed only for particular models and matching is determined by the order of approximation to the integrated probability. The selection of a matching prior is usually accompanied by a discussion of the degree of matching, with various definitions of matching available (e.g. Reid et al. 2003). Although matching is usually considered between cumulative probabilities, we match on the expected model parameter covariance. This second moment is commonly used as the basis for model parameter confidence intervals in physics, and can be broadly interpreted as fixing the multidimensional 'width' of a distribution.

Matching priors are candidates for non-informative priors in Bayesian inference, in that it is often assumed (explicitly or not) that the frequentist-style determination of confidence intervals incorporates no information from a prior. Really, there is simply no such thing as a non-informative prior. The frequentist philosophy is different from the Bayesian approach and provides different guarantees across notionally repeated experiments. However, given that the concept of 'errors' is often interpreted according to the frequentist philosophy, we think there is merit in making the widths of the errors consistent.

Matching priors (and frequentist analyses) violate the Likelihood Principle by using priors that vary with the sampling distribution of the experiment to be performed and the dimension of the model parameter space on to which the data distribution is projected. However, in general they only rely on the performance characteristics of that distribution under repeated sampling, as a way to 'break the tie' among a choice of prior distributions, in order to draw an inference. Thus, while there is debate about their validity and usage, it is clear that there are situations where they are useful.

In this paper, we argue that the analyses presented in Hartlap et al. (2007) and Dodelson & Schneider (2013) provide a method for calculating frequentist based confidence intervals for model parameters, and we show that these can be matched to credible intervals obtained from a Bayesian analysis as advocated by Sellentin & Heavens (2016). A similar calculation was performed by Percival et al. (2014) but we now use the methodology and resulting form for the posterior adopted by Sellentin & Heavens (2016), albeit using a different prior. This demonstrates how these different methods are related and the different assumptions being (sometimes implicitly) made when adopting one of these procedures for determining and quoting the coverage probability associated with an interval. The frequentist matched credible intervals are larger than those from Bayesian analyses with previously used priors, and hence this matching can also be considered conservative for inferences made from experiments.

The layout of our paper is as follows: Section 2 introduces the Bayesian problem that we want to solve, and considers how the posterior depends on the prior chosen, extending the Sellentin & Heavens (2016) approach to more general priors. Section 3 considers probabilities under the posterior and relates them to the distribution of the truth after repeated trials, allowing us to define a frequentist matching prior in Section 4. Section 5 demonstrates this approach using the simple problem of fitting a mean to correlated data, using both analytic derivations and Monte Carlo simulations. We apply our approach to a realistic cosmological analysis in Section 6, fitting mock tomographic cosmic shear data vector including auto- and cross-correlations matching that expected from the 5-year data of the Dark Energy Survey, demonstrating that this works well in a practical test, providing Bayesian credible intervals on model parameters that match the expected frequentist confidence intervals. We summarize our proposed method in Section 7, and conclude in Section 8.

2 CHOICE OF PRIOR TO USE IN A MODEL FIT

In this section we consider a full Bayesian analysis of the problem, considering different choices for the prior.

2.1 Posterior with an independence Jeffreys prior

The uninformative nature of the independence Jeffreys prior in general was introduced at the very start of Bayesian statistics (Jeffreys 1939) and is discussed in this specific situation in Sun & Berger (2006). It assumes for Gaussian data a uniform prior for the means, and a Jeffreys prior for the covariance matrix with means given (Berger & Sun 2008). The derivation of the posterior using this choice of prior, and application to astronomical situations was presented in Sellentin & Heavens (2016).

We assume the independence Jeffreys joint prior on the expectation value of the data and its covariance matrix given by equation (7). To calculate the required posterior using equation (3), we first note that S follows a Wishart distribution, f_W, and we can write

$$f(\Sigma|S) \propto f_W(S|\Sigma/(n_s-1), n_s-1)\, f(\boldsymbol{\mu}, \Sigma), \tag{9}$$

$$\propto |\Sigma|^{-\frac{n_s+n_d}{2}} \exp\left[-\frac{n_s-1}{2} Tr(\Sigma^{-1}S)\right], \tag{10}$$

$$\propto f_{W^{-1}}(\Sigma|(n_s-1)S, n_s-1), \tag{11}$$

which shows how, with this prior, the posterior for Σ has an inverse Wishart distribution, $f_{W^{-1}}$. The definitions of the multivariate distributions used in our work are included in Appendix A.

We now multiply by the Gaussian likelihood $f_N(\boldsymbol{x}_0|\boldsymbol{\mu}, \Sigma)$, which is simplest to consider in the form given in Appendix A, and integrate over Σ to find that

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) \propto \int \mathrm{d}\Sigma\, |\Sigma|^{-\frac{n_s+n_d+1}{2}} \exp\left[-\frac{1}{2} Tr(\Sigma^{-1}Q)\right], \tag{12}$$

where

$$Q = (n_s-1)S + (\boldsymbol{x}_0-\boldsymbol{\mu})(\boldsymbol{x}_0-\boldsymbol{\mu})^T. \tag{13}$$

This is an integral over the unnormalized inverse Wishart distribution (with parameter n_s), so we can read off the result from the normalization constant in equation (A2).

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) \propto |Q|^{-\frac{n_s}{2}}. \tag{14}$$

Comparing with the form of the multivariate t-distribution in equation (A4), we see that

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) = f_{t,n_s-n_d}\left(\boldsymbol{\mu}\,\middle|\,\boldsymbol{x}_0, \frac{n_s-1}{n_s-n_d}S\right), \tag{15}$$

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

which has mean $\boldsymbol{x}_0$ and covariance

$$\langle(\boldsymbol{\mu}-\boldsymbol{x}_0)(\boldsymbol{\mu}-\boldsymbol{x}_0)^T\rangle = \frac{n_s-1}{n_s-n_d-2}S = hS. \tag{16}$$

The use of the multivariate t-distribution as a replacement for the Gaussian assumption is often advocated on the grounds of robustness to outliers (Lange, Little & Taylor 1989), with the parameter ν, which in our context is $n_s - n_d$ used as a robustness tuning factor. In this section we have shown how it also arises when the covariance matrix is itself a random variable. It is also interesting to see that, with an independence Jeffreys prior, the Hartlap factor emerges in the recovered covariance, which could be considered natural given that using this prior brings in no further information on the posterior, and the inclusion of the Hartlap factor in some sense unbiases the posterior covariance. However, inferences made from the posterior about the covariance on model parameters are biased by the inclusion of this factor - while it unbiases the posterior against repeated trials of S, inferences about model parameter covariances made from the posterior are biased - and so it is not clear that this is what we actually want (see Section 8 for further discussion of this). We also note that a Gaussian posterior with a Hartlap correction yields a posterior covariance that agrees with that derived here, but has tail probabilities that are lower than the t-distribution, and may be in considerable error when data sets in tension are discussed and compared (see Appendix D).

In the next section we see that the multivariate t-distribution form for the posterior follows from any prior that is a power-law in $|\Sigma|$, and that the exponent of the power-law affects the recovered credible intervals.

2.2 Posterior with a general power-law prior

Let us now consider a more general joint prior on the mean and covariance matrix

$$f(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-\frac{m-n_s+n_d+1}{2}}. \tag{17}$$

The independence Jeffreys prior of Sellentin & Heavens (2016) corresponds to $m = n_s$. Both priors are uniform in the mean, which makes sense for a location parameter. The exact linear form for the exponent is chosen to simplify the downstream analysis, but is not important. It changes our conditional likelihood

$$f(\Sigma|S) \propto |\Sigma|^{-\frac{m+n_d}{2}} \exp\left[-\frac{n_s-1}{2}Tr(\Sigma^{-1}S)\right], \tag{18}$$

and we now have that

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) \propto \int d\Sigma|\Sigma|^{-\frac{m+n_d+1}{2}} \exp\left[-\frac{1}{2}Tr(\Sigma^{-1}Q)\right], \tag{19}$$

where Q is given by equation (13). The form of this equation still matches that of an unnormalized inverse Wishart distribution, but with different parameters, so we now have

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) \propto |Q|^{-\frac{m}{2}}. \tag{20}$$

Following through the derivation,

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) = f_{t,m-n_d}\left(\boldsymbol{\mu}\,\middle|\,\boldsymbol{x}_0, \frac{n_s-1}{m-n_d}S\right). \tag{21}$$

From the known properties of the multivariate t-distribution, this has mean $\boldsymbol{x}_0$ and covariance

$$\langle(\boldsymbol{\mu}-\boldsymbol{x}_0)(\boldsymbol{\mu}-\boldsymbol{x}_0)^T\rangle = \frac{n_s-1}{m-n_d-2}S. \tag{22}$$

As expected, setting $m = n_s$ gets us back to equation (16), and an expected covariance of hS. The covariance recovered from the distribution is directly related to the prior through m - as is natural in a Bayesian analysis.

3 MODEL PARAMETER COVARIANCES FROM POSTERIORS AND FROM THE PARAMETER DISTRIBUTION

We now consider different methods for characterizing our uncertainty about model parameters by comparing the model parameter covariances calculated using different assumptions.

Given a set of data ($\boldsymbol{x}_0$, S) and a prior parametrized by m, we first determine the Fisher matrix (Section 3.1) and then consider the model parameter covariance derived by computing probabilities under the posterior (Section 3.2). In order to construct a matching prior, for probabilities estimated using the Fisher matrix and probabilities calculated under the posterior, we need to determine the frequentist coverage probability that can be associated with the derived credible intervals. Formally, the coverage probability is a property of the procedure for constructing frequentist confidence intervals, and gives the proportion of repeated trials for which the interval contains the true value of interest. As we want to be able to interpret x per cent credible intervals as x per cent confidence intervals, we need to calculate the average size of the credible intervals of fixed probability over repeated trials. Finding the prior for which this is equal to the probability of finding the truth within each interval after repeated trials would then mean that we could interpret Bayesian credible intervals containing a particular probability with the same coverage probability. For simplicity, we work with the covariance rather than the intervals directly and hence we wish to know the average model parameter covariance recovered from the Fisher matrix or the posterior over repeated trials. For this, the multivariate t-distribution posterior has some differences from the expectation for a Gaussian posterior because the covariance of the posterior around the MAP model parameters depends on $\boldsymbol{x}_0$ in addition to S. Consequently, the distribution assumed for the data is important as we demonstrate by contrasting results assuming the data is drawn from a t-distribution, or from a Gaussian as is correct for our problem. The dependence of the model parameter covariance on $\boldsymbol{x}_0$ also affects data compression as we show in Appendix C.

We contrast the covariance estimated by integrating under the posterior with that calculated for the distribution of MAP solutions given the truth in Section 3.3, formally showing that, for our problem, they are very different for most choices of prior. In Section 4 we present the prior that matches these results.

3.1 Using the Fisher matrix

The Fisher information matrix (or simply the Fisher matrix), defined as

$$F(\boldsymbol{\theta})_{\alpha\beta} = E\left[\left(\frac{\partial}{\partial\theta_\alpha}\log f(\boldsymbol{x}_0|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_\beta}\log f(\boldsymbol{x}_0|\boldsymbol{\theta})\right)\right], \tag{23}$$

is a function of the likelihood. In Bayesian inference, the Bernstein–von Mises theorem provides the basis for using the Fisher matrix to provide confidence statements on parametric models, and the Cramér-Rao theorem shows that it forms a lower bound for the covariance of unbiased estimators of $\boldsymbol{\theta}$. In our case, we work from the posterior, as given in equation (21), and convert this to a likelihood assuming a uniform prior (albeit possibly improper) on the model parameters. Thus, in this section, we are not calculating the Fisher

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

matrix from the true likelihood of the data (remember that $\boldsymbol{x}_0$ are drawn from a Gaussian distribution with covariance Σ), but instead we use the Fisher matrix to estimate the expected information given the form of the posterior assumed.

We start by assuming that, around the peak of the posterior, we can define a patch of parameter space for which we can apply Bayes theorem to equation (21) with a uniform prior on $\boldsymbol{\mu}$. For this patch the likelihood for $\boldsymbol{x}_0$ is

$$f(\boldsymbol{x}_0|\boldsymbol{\mu}, S) = f_{t,m-n_d}\left(\boldsymbol{x}_0 \middle| \boldsymbol{\mu}, \frac{n_s - 1}{m - n_d} S\right). \tag{24}$$

The Fisher information matrix for the multivariate t-distribution with degrees of freedom ν and covariance Σ (Lange et al. 1989; Sellentin & Heavens 2017) is

$$F_t = \frac{\nu(\nu + n_d)}{(\nu - 2)(\nu + n_d + 2)} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}^T \Sigma^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}. \tag{25}$$

We see an extra term compared with the true Fisher Information matrix if the covariance matrix were known:

$$F_\Sigma = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}^T \Sigma^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}. \tag{26}$$

For completeness, the Gaussian Fisher Information matrix with covariance matrix S is

$$F_S = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}^T S^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}. \tag{27}$$

For the likelihood of equation (24), we have $\nu = m - n_d$ degrees of freedom and a covariance $(n_s - 1)S/(m - n_d - 2)$, so we have

$$F_t = \frac{m(m - n_d)}{(m + 2)(n_s - 1)} F_S. \tag{28}$$

This is the t-distribution Fisher matrix given the approximate scale matrix S.

As discussed at the start this section, we also want to determine the average credible interval that would be recovered given a set of realizations of S drawn from a Wishart distribution (i.e. by observers in parallel universes). To calculate this, we note that a property of the Wishart distribution is that for

$$f(S|\Sigma) = f_W(S|\Sigma/(n_s - 1), n_s - 1), \tag{29}$$

and M a $n_\theta \times n_d$ matrix, then

$$\begin{aligned} &f((MS^{-1}M^T)^{-1}|\Sigma) \\ &\quad = f_W\left((MS^{-1}M^T)^{-1} \middle| \frac{(M\Sigma^{-1}M^T)^{-1}}{n_s - 1}, n_s - n_d + n_\theta - 1\right), \end{aligned} \tag{30}$$

(see theorem 3.2.11 of Muirhead 1982). Thus, from equation (27), and using the mean of the Wishart distribution, we have that

$$\left\langle F_S^{-1} \right\rangle_S = \frac{n_s - n_d + n_\theta - 1}{n_s - 1} F_\Sigma^{-1}. \tag{31}$$

This equation can also be approximated by writing $(hS)^{-1}$ as a perturbation around Σ^{-1} and considering the second-order terms, as discussed in Appendix B, and used in Percival et al. (2014).

For the t-distribution Fisher matrix, from equation (28), we have that

$$\langle F_t^{-1} \rangle_S = \frac{(m + 2)(n_s - n_d + n_\theta - 1)}{m(m - n_d)} F_\Sigma^{-1}. \tag{32}$$

This shows that the error in the covariance matrix has an additional effect on the average model parameter credible intervals derived from a set of realizations of the scale matrix.

3.2 Computing probabilities under the posterior

We now consider credible intervals derived by computing probabilities under the posterior, based on the second moment of the distribution. While the Fisher matrix gives the form of the likelihood around the expected value, calculating probabilities under the posterior is the more common approach used for model parameter credible interval determination. We consider the case where we have a linear model with $\boldsymbol{\mu} = E\boldsymbol{\theta}$, for some generally non-square matrix E. Using equation (21) the posterior can be written

$$f(\boldsymbol{\theta}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{1}{n_s - 1}(\boldsymbol{x}_0 - E\boldsymbol{\theta})^T S^{-1} (\boldsymbol{x}_0 - E\boldsymbol{\theta})\right]^{-\frac{m}{2}}. \tag{33}$$

This can be manipulated to describe the posterior as a distribution around the MAP estimate. For a simple example of this for a Gaussian posterior, and a single-parameter model - fitting the mean to data - see Appendix E1. The same derivation can be seen in Appendix E2 for the case of fitting the mean using a t-distribution posterior. Keeping to a more general linear model, expanding the distribution, we have

$$\begin{aligned} &f(\boldsymbol{\theta}|\boldsymbol{x}_0, S) \\ &\quad \propto \left[1 + \frac{\boldsymbol{x}_0^T S^{-1}\boldsymbol{x}_0 - 2\boldsymbol{\theta}^T E^T S^{-1}\boldsymbol{x}_0 + \boldsymbol{\theta}^T E^T S^{-1} E\boldsymbol{\theta}}{n_s - 1}\right]^{-\frac{m}{2}}, \end{aligned} \tag{34}$$

using the symmetry of S^{-1} to simplify the cross terms. Setting $F_S = E^T S^{-1} E$ and $\boldsymbol{g} = E^T S^{-1}\boldsymbol{x}_0$ gives

$$\begin{aligned} &f(\boldsymbol{\theta}|\boldsymbol{x}_0, S) \\ &\quad \propto \left[1 + \frac{\boldsymbol{x}_0^T S^{-1}\boldsymbol{x}_0 - \boldsymbol{g}^T F_S^{-1}\boldsymbol{g} + \left(\boldsymbol{\theta} - F_S^{-1}\boldsymbol{g}\right)^T F_S \left(\boldsymbol{\theta} - F_S^{-1}\boldsymbol{g}\right)}{n_s - 1}\right]^{-\frac{m}{2}}. \end{aligned} \tag{35}$$

To finish the derivation, we need to complete the square, noting that if we now define

$$\boldsymbol{y} = \left(\boldsymbol{\theta} - F_S^{-1}\boldsymbol{g}\right)\left(\frac{n_s - 1}{m - n_\theta}\right)^{-\frac{1}{2}} \left[1 + \frac{\boldsymbol{x}_0^T S^{-1}\boldsymbol{x}_0 - \boldsymbol{g}^T F_S^{-1}\boldsymbol{g}}{n_s - 1}\right]^{-\frac{1}{2}}, \tag{36}$$

then the posterior reduces to the simple form

$$f(\boldsymbol{\theta}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{\boldsymbol{y}^T F_S^{-1}\boldsymbol{y}}{m - n_\theta}\right]^{-\frac{m}{2}}. \tag{37}$$

This shows that $\boldsymbol{y}$ is distributed with a multivariate t-distribution with $m - n_\theta$ degrees of freedom, such that the mean $\langle \boldsymbol{y} \rangle = \boldsymbol{0}$, and covariance $\langle \boldsymbol{y}\boldsymbol{y}^T \rangle = (m - n_\theta)F_S^{-1}/(m - n_\theta - 2)$.

We can write $\boldsymbol{\theta}$ in the form $\boldsymbol{\theta} = a\boldsymbol{y} + b$, which has the property that $\langle \boldsymbol{\theta} \rangle = a\langle \boldsymbol{y} \rangle + b$, and $\langle (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \rangle = a^2 \langle \boldsymbol{y}\boldsymbol{y}^T \rangle$. From this, we see that the distribution of $\boldsymbol{\theta}$ has mean $\hat{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle = F_S^{-1}\boldsymbol{g}$. The covariance of $\boldsymbol{\theta}$ around this for any value of $\boldsymbol{x}_0$ and S is

$$\begin{aligned} \langle (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \rangle &= \frac{n_s - 1}{m - n_\theta - 2} F_S^{-1} \\ &\quad \times \left[1 + \frac{\boldsymbol{x}_0^T S^{-1}\boldsymbol{x}_0 - \boldsymbol{g}^T F_S^{-1}\boldsymbol{g}}{n_s - 1}\right]. \end{aligned} \tag{38}$$

For a linear model, this expression can be used instead of integrating under the posterior for any realization of the data $(\boldsymbol{x}_0, S)$. Crucially, unlike the equivalent calculation for the Gaussian distribution (see Appendix E1 for this calculation in the special case of fitting the mean to data), the model parameter covariance depends on the value

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

of $\boldsymbol{x}_0$. Thus, the size of the credible intervals we derive from our fit will change if we change the data.

We now consider the model parameter covariance recovered by integrating under the posterior, averaged over a set of values of $\boldsymbol{x}_0$ and S. We start by considering $\boldsymbol{x}_0$ distributed according to the t-distribution, and a Wishart distributed S. However, while we adopt a posterior that has multivariate t-distribution form, the data itself are actually Gaussian distributed with covariance Σ, and so we consider this case afterwards.

3.2.1 *Data distributed according to the t-distribution*

We can now calculate the expected covariance recovered for the model parameters, averaging over multiple realizations of the data $(\boldsymbol{x}_0, S)$. We start by assuming that the same covariance matrix approximation S is used for all realizations. In this case, F_S^{-1} is fixed, and we need to replace the terms $\boldsymbol{x}_0^T S^{-1} \boldsymbol{x}_0$ and $\boldsymbol{g}^T F_S^{-1} \boldsymbol{g}$ by the relevant expected values. To calculate these, we make use of the fact that we have set up the problem such that $\langle \boldsymbol{x}_0 \rangle$ is the zero vector, and make use of the identity $\boldsymbol{x}_0^T S^{-1} \boldsymbol{x}_0 = Tr(S^{-1} \boldsymbol{x}_0 \boldsymbol{x}_0^T)$. We find that, for a set of data drawn from a multivariate t-distribution as in equation (24), we have

$$\left\langle \boldsymbol{x}_0^T S^{-1} \boldsymbol{x}_0 \right\rangle_x = \frac{n_s - 1}{m - n_d - 2} n_d, \tag{39}$$

$$\left\langle \boldsymbol{g}^T F_S^{-1} \boldsymbol{g} \right\rangle_x = \frac{n_s - 1}{m - n_d - 2} n_\theta. \tag{40}$$

Putting these values in to equation (38), the covariance for $\boldsymbol{\theta}$ reduces to

$$\langle (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \rangle_x = \frac{n_s - 1}{m - n_d - 2} F_S^{-1}. \tag{41}$$

The expectation over multiple S matrices drawn from a Wishart distribution can easily be calculated using equation (31),

$$\langle (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \rangle_{x,S} = \frac{n_s - n_d + n_\theta - 1}{m - n_d - 2} F_\Sigma^{-1}. \tag{42}$$

3.2.2 *Gaussian distributed data*

For a set of data drawn from a Gaussian distribution with covariance Σ, we have

$$\left\langle \boldsymbol{x}_0^T S^{-1} \boldsymbol{x}_0 \right\rangle_x = Tr[S^{-1}\Sigma], \tag{43}$$

$$\left\langle \boldsymbol{g}^T F_S^{-1} \boldsymbol{g} \right\rangle_x = Tr[F_S^{-1} E^T S^{-1} \Sigma S^{-1} E]. \tag{44}$$

To go one step further and consider the expected model parameter covariance allowing for multiple S matrices drawn from a Wishart distribution, we now need to find expressions for the expectation of all of the terms in equation (38). We have equation (31) for $\langle F_S^{-1} \rangle_S$, and

$$\left\langle F_S^{-1} Tr[S^{-1}\Sigma] \right\rangle_S \simeq [n_d + B(n_d(n_\theta + 1) - 2)] F_\Sigma^{-1}, \tag{45}$$

$$\left\langle F_S^{-1} Tr[F_S^{-1} E^T S^{-1} \Sigma S^{-1} E] \right\rangle_S \simeq [n_\theta + B(n_\theta(n_d + 1) - 2)] F_\Sigma^{-1}, \tag{46}$$

where B is given in equation (B2). To get these expressions, we have used the perturbative expressions as described in Appendix B.

The end result is that we should expect the average model parameter covariance recovered integrating under the posterior after repeated trials where the data is drawn from a Gaussian distribution with true covariance Σ, and S is drawn from a Wishart distribution to be

$$\langle (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \rangle_{x,S} \simeq \frac{n_s - 1 + B(n_d - n_\theta)}{m - n_\theta - 2} F_\Sigma^{-1}, \tag{47}$$

to second order. The difference between this expression and that of equation (42) shows the importance of the distribution of $\boldsymbol{x}_0$ in calculating the average model parameter covariance recovered. The situation with Gaussian distributed data matches the set-up of our problem: that of considering observers in multiple universes.

3.3 The distribution of the difference between MAP estimate and the truth

We now contrast these estimates of the model parameter covariance against the distribution of recovered maximum a posteriori model parameter values recovered from reruns of the experiment being performed. A linear model is assumed, so we have the symmetry that the distribution of MAP solutions about the truth is the same as the distribution of the truth around a particular MAP solution (when the truth is sampled from a uniform prior). By comparing the results in Section 3.2 to those from a Gaussian posterior, we see that the MAP estimate for the model parameters is the same whether using a Gaussian or t-distribution posterior and so we do not need to distinguish between these choices.

We therefore start assuming a Gaussian posterior distribution as in Dodelson & Schneider (2013). As discussed in Section 3.2, the MAP estimate for a linear model can be written

$$\hat{\boldsymbol{\theta}} = F_S^{-1} \boldsymbol{g} = F_S^{-1} E^T S^{-1} \boldsymbol{x}_0, \tag{48}$$

which can also be recovered as the first-order solution for more general models by Taylor expanding the posterior around the MAP estimates of the model parameters. Here we have assumed, without loss of generality, that the true values are $\hat{\boldsymbol{\theta}} = \boldsymbol{0}$.

We can now obtain an estimate of the scatter on model parameters provided by different experiments, where we consider different $\boldsymbol{x}_0$ drawn from a Gaussian distribution, and S from a Wishart distribution given the true model $\langle \hat{\boldsymbol{\theta}}^T \hat{\boldsymbol{\theta}} \rangle_{x,S}$. To do this, we use the fact that $\langle \boldsymbol{x}_0 \boldsymbol{x}_0^T \rangle_x = \Sigma$, so that

$$\langle \hat{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}^T \rangle_x = \left\langle F_S^{-1} E^T S^{-1} \Sigma S^{-1} E F_S^{-1} \right\rangle. \tag{49}$$

This can be solved to second order, using the expression in equation (B1), considering an expansion of $(hS)^{-1}$ around Σ^{-1}. As described in Appendix B, the second-order solution is

$$\langle \hat{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}^T \rangle_{x,S} \simeq [1 + B(n_d - n_\theta)] F_\Sigma^{-1}, \tag{50}$$

which is the distribution of MAP estimates made from a set of simulations that is independent of those used to estimate the covariance matrix S. This was the primary result of Dodelson & Schneider (2013). Because we assume a linear model, this model parameter covariance is also that of the distribution of the truth around the MAP solution, assuming a uniform prior on the model parameters. It is therefore the covariance of the distribution from which frequentist confidence intervals on model parameters are derived.

4 FREQUENTIST MATCHING PRIOR

We now consider how to derive a matching prior that will allow the average model parameter covariance derived from the Bayesian analysis described above to match the recovered covariance of the truth around the MAP estimate. To do this, we compare and match

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

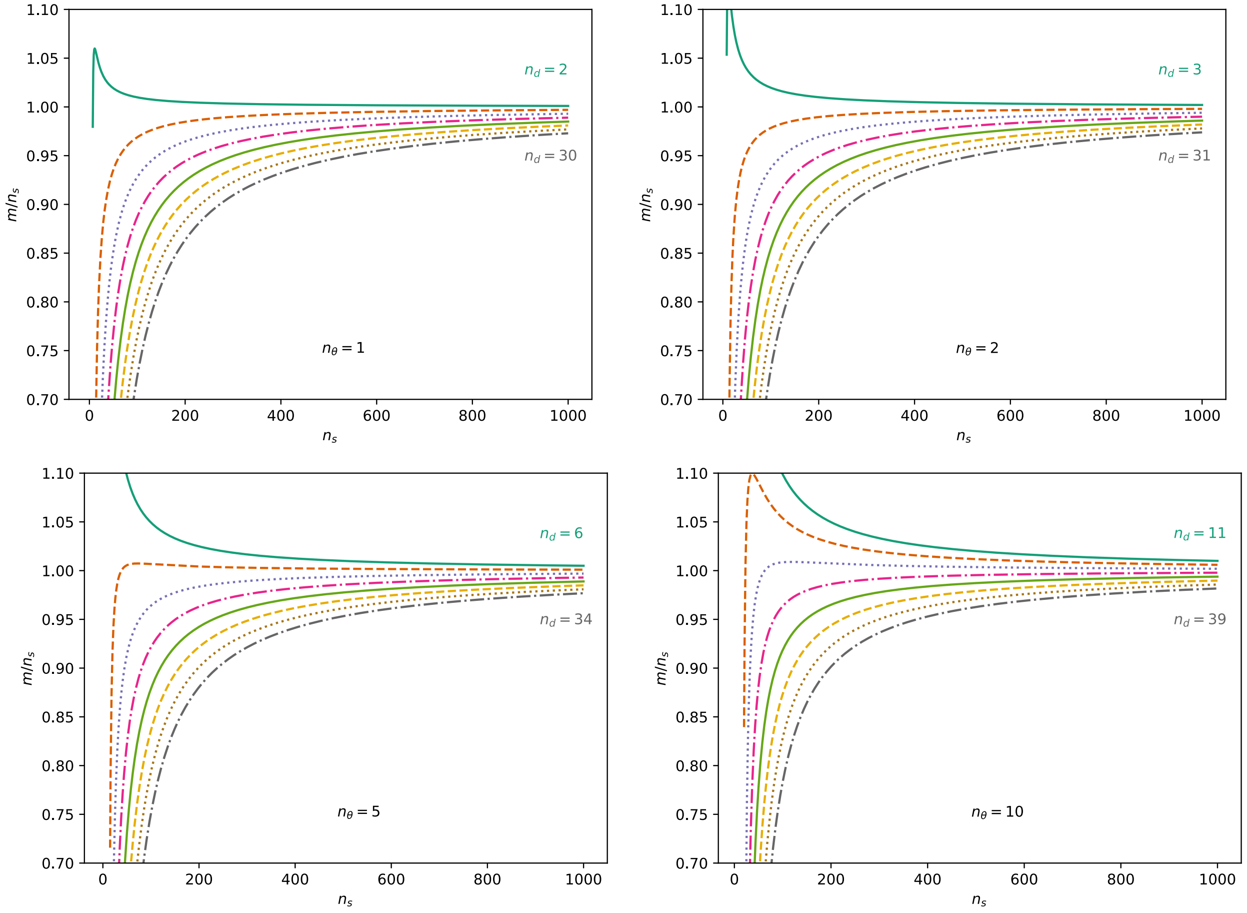

Figure 1. Variation of the power-law exponent m_{match} required for a prior that, on average over repeated trials gives a posterior with covariance that matches that expected for the distribution of MAP values (solid lines as given in equation (51)). We show how m_{match}/n_s varies with n_s (x-axis), n_d (different lines), and n_θ (different panels).

equations (47) and (50) to derive a Bayesian posterior parametrized by m_{match} that gives a posterior distribution that, averaged over multiple trials, has a model parameter covariance that matches the distribution of MAP estimates that we would get from repeating the experiment. This assumes that, for these repeated trials, $\boldsymbol{x}_0$ is drawn from a Gaussian distribution around the true cosmological model. In this case, the equation for m_{match} is

$$m_{\mathrm{match}} = n_\theta + 2 + \frac{n_s - 1 + B(n_d - n_\theta)}{1 + B(n_d - n_\theta)}. \tag{51}$$

The resulting values of m_{match} are compared in Fig. 1 for a range of values of n_s, n_d, and n_θ. As can be seen, m_{match} tends towards the Sellentin & Heavens (2016) solution $m = n_s$ for large values of n_s. However, there are differences, especially when $n_s \sim n_d$ and the posterior is more influenced by the prior than when many more simulations are available. We note that this is derived under a number of assumptions, particularly that of a linear model, and so this is still an approximation to a true matched posterior given a more complicated shape and non-linear model dependence. In particular, we caution that the moment-matching prior is not invariant to reparametrization. We find that the exponent for $n_d = 2$, $n_\theta = 1$ is very close to that derived from the right-Haar prior (based on Cholesky decomposition of the covariance matrix), which has some exact matching properties for Gaussian variables (Sun & Berger 2006).

5 TESTING WITH A SIMPLE MEAN FITTING MODEL

The resulting covariance matrix for the model parameters is tested and explored by considering a simple model - that of fitting a mean value to correlated data. We create Monte Carlo simulations that step through different realizations of the data (Gaussian distributed with covariance Σ, chosen for convenience to be the identity matrix) and analysed with covariance matrix S drawn from a Wishart distribution (degrees of freedom $n_s - 1$ and scale matrix Σ). Inferences are made about credible intervals assuming different choices for the posterior, and the derived estimates of the model parameter covariances are then averaged over multiple realizations. Averaging over realizations of the data and covariance matrix S in this way most naturally follows the ethos behind the derivation in Section 3.2. We also record the MAP estimates for the model, and consider the distribution of these MAP estimates around the true values and measure the variance of this distribution.

We create large numbers of realizations of data $\boldsymbol{x}_0$ and covariance matrices S and then fit to each assuming different expressions for the posterior. For each covariance matrix S, we create n_s different versions of $\boldsymbol{x}_0$, and we create 100 000 different covariance matrices. To speed up these calculations we use analytic marginalization over the posterior for each S, as outlined in Appendix E, rather than numerically integrating under the posterior for each, and use library routines to calculate realizations of Wishart matrices. We still use

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

Monte Carlo results for different values of S, and the distribution of MAP parameters. Results are shown in Fig. 2, which shows that, as expected, we can choose a prior to match the model parameter covariance recovered from the posterior to that calculated in a frequentist style approach where we look at the spread of recovered MAP estimates. Given that this derivation best matched the set-up of the Monte Carlo simulations, using $m_{\rm match}$ provides an excellent fit to the numerical results.

6 TESTING AGAINST A NONLINEAR MODEL

To test the performance of the different posterior distributions discussed in Sections 2 and 4 in a realistic cosmological setting we adopt a mock experiment as also considered by Friedrich & Eifler (2018). They simulated a tomographic cosmic shear data vector including auto- and cross-correlations of $\xi_\pm$ in five source redshift bins on a survey area of 5000 deg^2 (hence mimicking 5-yr data of the Dark Energy Survey, cf. their table 1 for details). Overall this data vector contains 450 data points. Around a true data vector computed at a cosmology with $(\Omega_m, \sigma_8, w_0) = (0.3156, 0.831, -1)$ we draw 1000 Gaussian random realizations assuming a theoretical covariance matrix derived using the halo model to describe non-linear clustering. Here Ω_m is the present-day cosmological matter density, σ_8 is the rms density fluctuations in spheres of radius 8 h^{-1}Mpc, and w_0 is the Dark Energy equation-of-state parameter. Both the covariance calculation and subsequent analyses of the mock data vectors are carried out with the CosmoLike toolkit (Krause & Eifler 2017).

In Fig. 3 we show marginalized posterior constraints in the Ω_m–σ_8 plane obtained from the first three of our random realizations using different posteriors. The grey shaded contours were obtained using the true analytic covariance that was also used to draw our mock data vectors. The orange contours assume that there is a covariance estimate from 650 simulations (i.e. 200 more than data points) and that this estimate is used in the posterior of Sellentin & Heavens (2016) to obtain the constraints (we draw a new covariance estimate for each data vector from a Wishart distribution). Note that all contours within each individual panel of Fig. 3 are derived from the same data vector realizations. Despite that, there is a noticeable additional scatter between the two sets of contours - this is exactly the effect of additional scatter of MAP estimates due to noisy covariance estimates described by Dodelson & Schneider (2013). The blue contours are the modified version of the posterior with a $m_{\rm match}$ prior chosen to match this additional scatter.

To assess the performance of our matched prior more quantitatively we run our Markov Chain Monte Carlo routine to explore the posteriors around all 1000 random realizations of our data vector. Fig. 4 compares how often the true cosmology underlying our numerical experiment is located inside the 68 per cent (left-hand panel) and 95 per cent (right-hand panel) confidence regions of the full three-dimensional parameter space when using different covariance matrices and different posterior distributions. Here we are considering the credible intervals derived from our Bayesian analysis work as frequentist confidence intervals. The grey band in each panel assumes that the true covariance is known. The green crosses represent the commonly used approach of a Gaussian likelihood with Hartlap corrected precision matrix as estimated from different numbers of simulations (x-axis in both panels). The orange dots use the independence Jeffreys prior advocated by Sellentin & Heavens (2016) and the resulting t-distribution instead of the Hartlap-corrected Gaussian likelihood. The blue triangles show the coverage achieved with a matched prior that uses equation (51) to compute

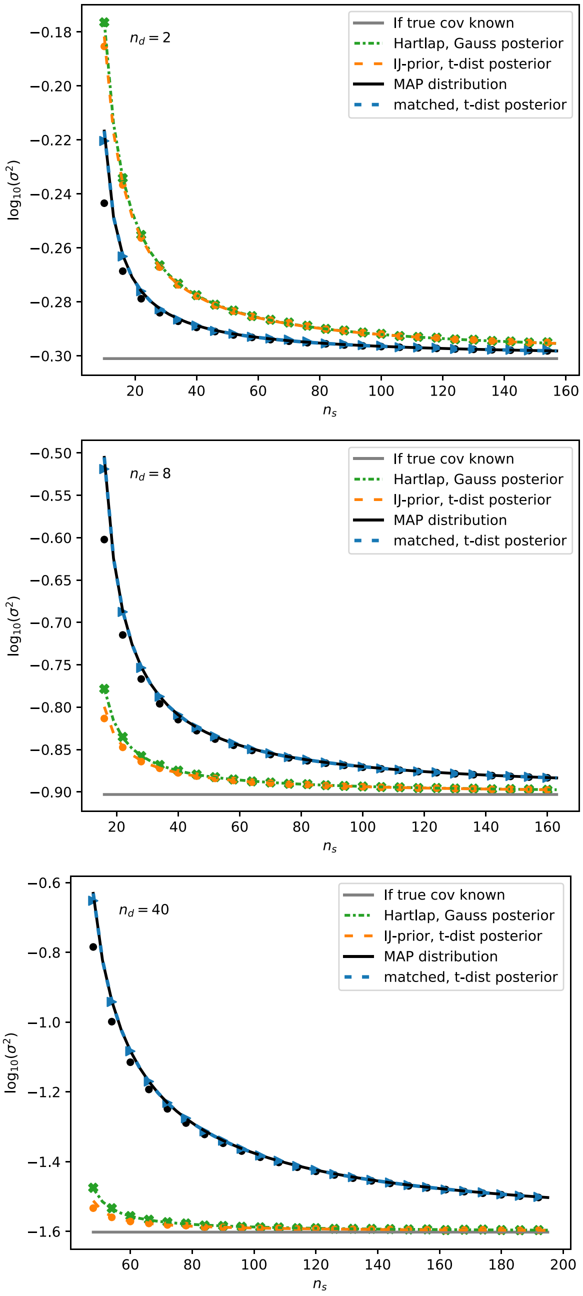

Figure 2. Average recovered model parameter variance calculated in different ways when fitting the mean $\bar{\mu}$, to a set of correlated Gaussian data. The grey solid line shows the result that would have been obtained from the posterior if we had known the data covariance matrix perfectly, taking the confidence interval as the root of the variance. The black solid line (model from Dodelson & Schneider 2013) and solid black points (Monte Carlo measurements) show the root of the variance determined from the distribution of recovered values. The difference is likely due to the perturbative nature of the derivation of the expectation. Green and orange lines and symbols show that we recover the similar distributions from both Hartlap-corrected Gaussian and t-distribution posteriors calculated with an independence Jeffreys prior as advocated in Sellentin & Heavens (2016). This can easily be understood as the posteriors have the same variance. The blue dashed lines and triangles show the result of using a t-distribution posterior with $m_{\rm match}$, corresponding to equation (51). As can be seen, this prior is able to match the variance recovered by integrating under the posterior averaged over our realizations, with the scatter recovered from MAP estimates.

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

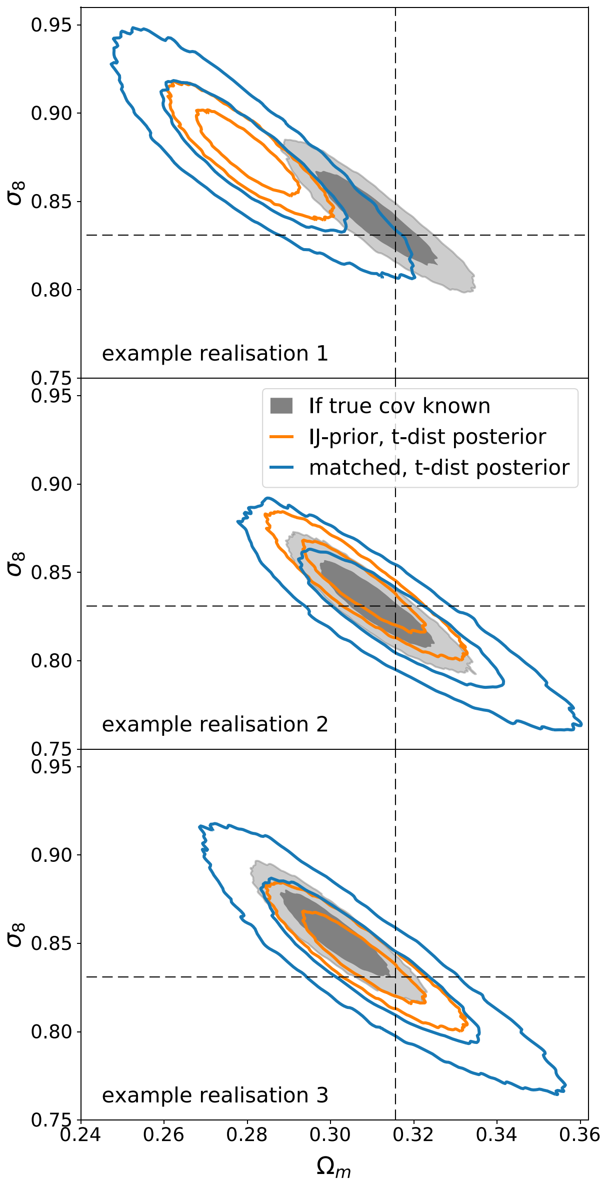

Figure 3. Contours containing 68 per cent and 95 per cent probability, marginalizing under the posterior in the Ω_m–σ_8 plane, obtained from realizations of DES-like weak lensing data vectors. Each panel is for a different random set of data $\boldsymbol{x}_0$ and covariance S drawn from Gaussian and Wishart distributions respectively. The relevant parameters of this run for the posterior are $n_s = 650$, $n_d = 450$, and $n_\theta = 4$. Contours are shown calculated using the true covariance matrix with a Gaussian posterior (grey shading), and two versions of the t-distribution posterior, one with $m = n_s$ as derived using an independence Jeffreys prior for the true covariance as in Sellentin & Heavens (2016) (orange), and one using a covariance-matching prior derived for linear models (blue). The dashed lines mark the expected values of both parameters.

the exponent m. This likelihood indeed manages to achieve coverage factions of approximately 68 per cent and 95 per cent respectively. The red squares show the coverage obtained from simply re-scaling the Gaussian log-likelihood in the manner advocated by Percival et al. (2014), which is also close to 68 per cent and 95 per cent respectively. The dash–dotted line shows the coverage that is expected for the standard Gaussian likelihood based on the calculations of Dodelson & Schneider (2013).

7 SUMMARY

Our suggested way forward is quite simple - in situations where the covariance matrix S for Gaussian data is itself a random variable drawn from a Wishart distribution with $n_s - 1$ degrees of freedom, for example when it is constructed from n_s mock samples, then we propose a frequentist matching prior that is uniform in $\boldsymbol{\mu}$ and depends on Σ as $|\Sigma|^{-(m-n_s+n_d+1)/2}$, leading to a posterior

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{\chi^2}{(n_s - 1)}\right]^{-\frac{m}{2}}, \tag{52}$$

where

$$\chi^2 = (\boldsymbol{x}_0 - \boldsymbol{\mu})^T S^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}). \tag{53}$$

The power-law index m is given by equation (51), and repeated here for completeness

$$m = n_\theta + 2 + \frac{n_s - 1 + B(n_d - n_\theta)}{1 + B(n_d - n_\theta)}, \tag{54}$$

$$B = \frac{(n_s - n_d - 2)}{(n_s - n_d - 1)(n_s - n_d - 4)}, \tag{55}$$

where n_d is the number of data points and n_θ the number of parameters. This will lead to credible intervals that can also be interpreted as confidence intervals with approximately the same coverage probability. Note that this expression does not require any extra factors of h, or other terms – i.e. S is the approximate covariance matrix, and S^{-1} its inverse. This enables a Bayesian analysis, with a matching prior designed with this frequency-matching property. In general, this procedure increases the model parameter credible intervals compared with those derived from the more usual independence Jeffreys prior on the true data covariance, and therefore can be considered a more conservative choice for making deductions from data.

If the reader prefers to approximate the posterior using a Gaussian distribution, then rather than inverting S or hS, the matrix $(S^{'})^{-1}$ to be used when calculating χ^2 should be the inverse of

$$S' = \frac{(n_s - 1)[1 + B(n_d - n_\theta)]}{n_s - n_d + n_\theta - 1} S, \tag{56}$$

which matches the method proposed in Percival et al. (2014), replacing one of the approximations used there with an exact expression. To derive this, consider the factor by which we must multiply equation (31) to obtain equation (50) – matching the model parameter covariance expected from integrating under the posterior with that from the distribution of MAP solutions.

Both the Gaussian approximation and our preferred t-distribution solution give model parameter covariances that are very similar to the suggestion of Friedrich & Eifler (2018) when n_θ is small. They proposed multiplying the Sellentin & Heavens (2016) posterior by the Dodelson & Schneider (2013) factor of $1 + B(n_d - n_\theta)$. To see the empirical similarity, note that the Sellentin & Heavens (2016) posterior gives a covariance for the distribution of $\boldsymbol{\mu}$ around $\boldsymbol{x}_0$ of hS, and compare equation (56) to $hS \times [1 + B(n_d - n_\theta)]$.

8 CONCLUSIONS

The primary result in our paper is presented in Section 7, which provides a frequentist-matching prior: i.e. the exponent in a power-law prior on the determinant of the true data covariance matrix required to yield a posterior model parameter covariance matching the distribution of true parameter values with respect to maximum likelihood estimates (and vice-versa for the linear models we

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

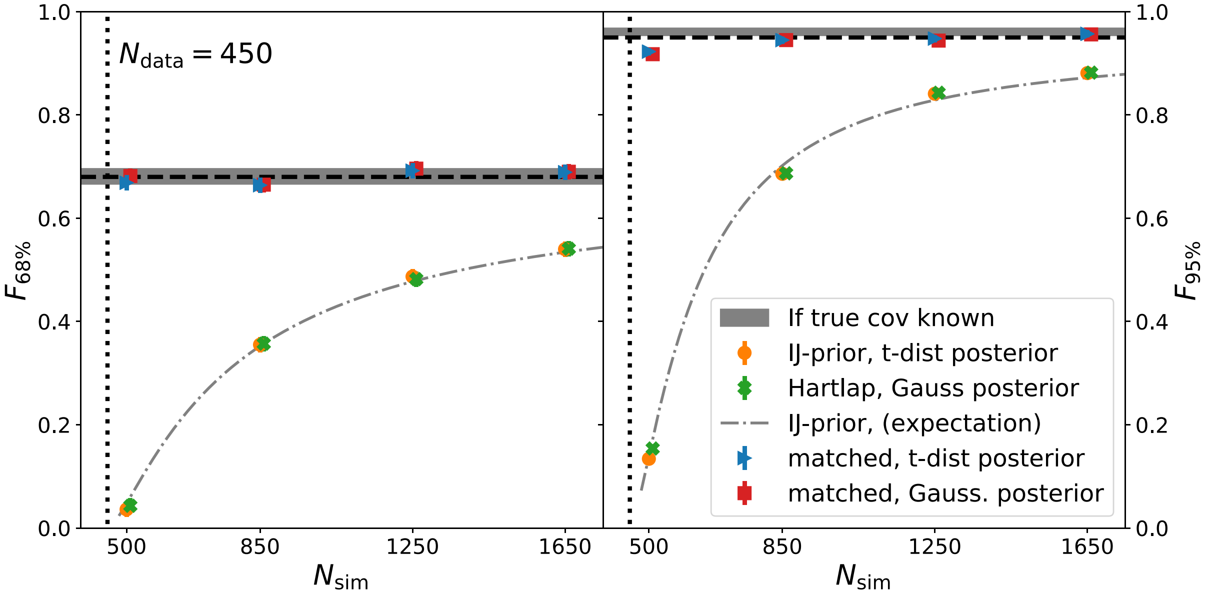

Figure 4. Comparing how often the true cosmology underlying our numerical experiment of Section 6 is found inside the 68 per cent (left panel) and 95 per cent (right panel) credible regions when using different covariance matrices and different posterior distributions. Note that we are comparing how well a credible region works as a confidence interval, and hence this test is not fair. The grey band in each panel assumes that the true covariance is known, in which case, with the problem being considered the credible interval also works as a confidence interval. The width of the band indicates the expected credible interval containing a coverage probability of 68 per cent from 1000 realizations, assuming a binomial distribution for the number of successes. Hence the horizontal black dashed line, which indicated the expected value does not have to lie in the middle of this band. The green crosses represent the common approach of a Gaussian likelihood with Hartlap corrected precision matrix for different numbers of simulations used to estimate the covariance matrix (x-axis in both panels). The vertical dotted line marks $n_s = n_d$. The orange dots use the independence Jeffreys prior advocated by Sellentin & Heavens (2016) and the resulting t-distribution instead of the Hartlap corrected Gaussian likelihood. The blue triangles show the coverage achieved with a matched prior that uses equation (51) to compute the exponent m, and the red squares show the coverage obtained from simply rescaling the Gaussian log-likelihood in the manner advocated by Percival et al. (2014), as in equation (56). The dash-dotted line show the coverage that is expected for the t-distribution posterior calculated using the independence Jeffreys prior.

consider). Our analysis lies at the interface between Bayesian and frequentist analyses: allowing an analysis that results in multiple interpretations of the same parameter intervals with the same probability. In order to derive this, we have assumed a linearized model, but have demonstrated broader applicability using a realistic non-linear model fit. Note that, in general, our results will not be valid for arbitrary non-linear models or reparametrizations. The use of this formalism for parameter inference when the covariance matrix is itself approximate offers a way to satisfy scientists whose intuition is based on frequentist style measures and those who wish for the analysis to be Bayesian in construct (which is often simpler for practical application).

We initially considered an independence Jeffreys prior on the true covariance matrix, as advocated in Sellentin & Heavens (2016). We showed that this leads to a posterior with covariance around the model parameters that matches that assuming a Gaussian posterior after scaling the data covariance matrix by the Hartlap factor. The derived model parameter covariance does not match that from the distribution of MAP estimates found by Dodelson & Schneider (2013), which is understandable given that they are calculating different distributions. We have considered alternative priors that are powers of the determinant of the true covariance matrix and which yield posteriors with frequentist coverage, at least at the level of covariance of the distributions. Using this allows the interpretation of credible intervals as confidence intervals with approximately the same probability. Because of the choice of a power-law prior, the posteriors of interest have the form of a multivariate t-distribution. For this form, the distribution of the posterior around the MAP estimate depends on the specific data realization - this can clearly be seen in equation (38). In comparison, for a Gaussian posterior, the distribution around the MAP estimates is independent of the data and depends only on the data covariance matrix S. This complicates the matching. We therefore consider the recovered model parameter covariance averaged over a set of data: here the distribution of that data matters. Formally, we calculate the frequentist coverage probability for a set of credible intervals, with a view to matching this probability to that from the distribution of MAP solutions.

Although we have a t-distribution posterior, the distribution of data is Gaussian, and so we cannot directly use either the t-distribution Fisher matrix (this led to expected covariance on model parameters as in equation (32)), or integrate under the posterior assuming the data is distributed according to a multivariate t-distribution (leading to equation (42)). Instead, we have to consider the Gaussian distribution of data when determining the average model parameter covariance that would be recovered from the posterior after repeated trials (giving equation (47)). We also note that this dependence on the data complicates data compression: the credible intervals recovered from compressed data do not necessarily match those recovered from the full data even for linear models where the compression is optimally performed to give the same MAP estimates (see Appendix C).

The prior that we advocate depends on the properties of the data and the problem, particularly n_s, n_d, and n_θ. Having priors that depend on the expected form of the posterior is quite common (although they should obviously not depend on the actual data observed), especially in the objective Bayesian approach (see Heavens & Sellentin 2018 for

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

an application to cosmology), so we do not see this as a fundamental problem, although it does conflict with the Bayesian notion of the prior as an expression of the state of knowledge before the experiment is performed.

One might also worry that our matching criterion is, in a sense, linking the posterior and properties of the data that depend on the likelihood. But the posterior should answer the question of what is the truth given the data, while the likelihood considers the data given the truth. These are fundamentally different things, and so why are we matching posterior and likelihood widths? If we compare the covariance inherent in the likelihood and the posterior for multivariate Gaussian distributions, then we might consider an approximate link where $\sigma^2_{\rm post} = \sigma^2_{\rm like} + \sigma^2_{\rm prior}$. This would be exact if all distributions were Gaussian, or we were working in the Gaussian limit. In this limit, the standard prior on the covariance used in the posterior directly adds to the covariance we assume for our experimental result. Translating through to model parameters, both contributions still contribute. So we see that the prior choice is related to the credible interval quoted for experimental measurements and forms the link between posterior and likelihood. A prior is chosen such that it does not change this covariance, and so in this sense our matching prior is an uninformative prior for the model parameters.

Using the multivariate t-distribution posterior makes the analysis attractive in a Bayesian sense, as it matches the problem with fewer approximations. In general, approximating the posterior as Gaussian has a relatively small effect on the posterior surface for 1σ and 2σ intervals, and in the examples we have considered less so than the choice of prior (see Appendix D). Even so, we recommend using the multivariate t-distribution with the revised prior as this represents a consistent Bayesian approach. Moreover, the tail probabilities can be much greater than those of the equivalent Gaussian, which can be in error when tensions between data sets are considered. In this case, we need to be careful about the interpretation of $N\sigma$ confidence intervals, as discussed in Appendix D. For those that cannot contemplate a posterior with a form other than Gaussian, we have included the alternative correction to use instead of the Hartlap factor for an approximate Gaussian posterior in Section 7.

When $n_\theta = n_d$, equation (51) gives that $m = n_s + n_d + 1$, and the prior reduces to $|\Sigma|^{-(n_d+1)}$. For this prior, the covariance of the posterior distribution as given in equation (22) reduces to S. From the properties of the Wishart distribution, this has expected value Σ matching the covariance of the frequentist distribution from which the data were assumed to be drawn. Note that no factor of h is required in the posterior, or in the Gaussian approximation to get this result. To understand why not, note that the rationale often used to justify using a Gaussian posterior based on a covariance hS (i.e. including a factor h) is that the inverse matrix S^{-1} is a biased estimate of Σ^{-1}, and this is corrected by using hS rather than S. Thus the argument goes that we should use hS in the posterior. However, we should consider that the model parameter covariance derived from the posterior is biased in the opposite way requiring an extra factor h^{-1} following the same rationale. To see this, consider equation (30), which shows that the model parameter covariance from a set of repeated trials each with a different S (with no h factor) is Wishart distributed with expectation given by a function of Σ. Where $n_\theta = n_d$, and we fit for the values of $\boldsymbol{\mu}$, the derived covariance reduces to S with expectation Σ, matching that we would expect given the Gaussian distribution of the data. Including h would have biased our errors compared to this expected value. Thus, explicitly including the Hartlap factor in a posterior to correct for a bias in S^{-1} is not just wrong from a Bayesian standpoint, but the standard rationale for its application misses a crucial step. Our proposed posterior consistently corrects for any potential biases due to having skewed distributions without any need for extra ad hoc factors.

Finally, we note that we form a matching prior based on the recovered model parameter covariance and not the distribution, as is more standard in statistical analyses. We do this because the covariance of the posterior distribution for model parameters offers a simple way to match the 'width' of two distributions, and that we can determine simple results for a power-law prior where we only have one degree of freedom and so only one degree of matching is possible. An extension to this work would be to consider varying the form of the prior beyond a simple power-law of the determinant of the true data covariance matrix to better match the shape of the posterior, in line with the more standard matching criterion used in statistics. We could also have directly compared credible intervals and confidence intervals - i.e. averaged over σ rather than the model parameter covariance where necessary, but we do not expect that this would change our results significantly compared with our chosen matching criterion based on covariance.

ACKNOWLEDGEMENTS

WJP acknowledges useful conversations with Michael Matesic and En Long regarding speeding up running the Monte Carlo realizations. We thank Daniel Farewell for helpful comments on an early version of the draft and Tim Eifler for useful comments on the draft and for providing the CosmoLike toolkit. We thank the referee, James Buchanan, for their careful review of the paper and for the corrections and suggestions provided.

Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities.

This research was enabled in part by support provided by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca).

DATA AVAILABILITY

No data were used in this paper, which are theoretical in nature.

REFERENCES

Alam S. et al., 2017, MNRAS, 470, 2617
Berger J., Sun D., 2008, Ann. Stat., 36, 963
Brouwer M. M. et al., 2018, MNRAS, 481, 5189
Chang T., Eaves D., 1990, Ann. Stat., 18, 1595
DES Collaboration, 2021, preprint (arXiv:2105.13549)
Dodelson S., Schneider M. D., 2013, Phys. Rev. D, 88, 063537
eBOSS Collaboration et al., 2021, Phys. Rev. D, 103, 083533
Friedrich O., Eifler T., 2018, MNRAS, 473, 4150
Friedrich O., Seitz S., Eifler T. F., Gruen D., 2016, MNRAS, 456, 2662
Friedrich O. et al., 2021, MNRAS, 508, 3125
Ghosh M., 2011, Statist. Sci., 26, 187
Giesser S., Cornfield J., 1963, J. R. Stat. Soc. B, 25, 368
Gruen D. et al., 2018, Phys. Rev. D, 98, 023507
Halder A., Friedrich O., Seitz S., Varga T. N., 2021, MNRAS, 506, 2780
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Heavens A. F., Sellentin E., 2018, J. Cosmol. Astropart. Phys., 2018, 047
Heymans C. et al., 2021, A&A, 646, A140
Jeffreys H., 1939, Theory of Probability. The Clarendon Press, Oxford
Joachimi B. et al., 2021, A&A, 646, A129
Kacprzak T. et al., 2016, MNRAS, 463, 3653

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

Kaufman G. M., 1967, Report No. 6710, Center for Operations Research and Econometrics. Catholic University of Louvain, Heverlee, Belgium
Krause E., Eifler T., 2017, MNRAS, 470, 2100
Lange K. L., Little R. J. A., Taylor J. M. G., 1989, J. Am. Stat. Assoc., 84, 881
Lindley D., 1958, J. R. Stat. Soc. B, 20, 102
Loredo T. J., 2012, in Joseph M. H.ed., Astrostatistical Challenges for the New Astronomy, Bayesian Astrostatistics: A Backward Look to the Future. Springer, New York, p. 15
Martinet N. et al., 2018, MNRAS, 474, 712
Muirhead R., 1982, Aspects of Multivariate Statistical Theory. Wiley, New Jersey
Norberg P., Baugh C. M., Gaztañaga E., Croton D. J., 2009, MNRAS, 396, 19
Percival W. J. et al., 2014, MNRAS, 439, 2531
Reid N., Mukerjee R., Fraser D. A. S., 2003, Lecture Notes-Monograph Series, 42, 31
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Sellentin E., Heavens A. F., 2017, MNRAS, 464, 4658
Sellentin E., Starck J.-L., 2019, J. Cosmol. Astropart. Phys., 2019, 021
Sun D., Berger J., 2006, Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics Benidorm. Alicante, Spain
Taylor A., Joachimi B., 2014, MNRAS, 442, 2728
Taylor A., Joachimi B., Kitching T., 2013, MNRAS, 432, 1928
Welch B. L., Peers H. W., 1963, On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 25. Wiley, Hoboken, NJ, p. 318

APPENDIX A: MULTIVARIATE DISTRIBUTIONS

Some multivariate distributions with data dimension n_d are listed here for reference:

The Wishart distribution

$$f_W(S|R,\nu) = \frac{|S|^{\frac{\nu-n_d-1}{2}}\exp\left[-\frac{1}{2}Tr(R^{-1}S)\right]}{2^{\frac{\nu n_d}{2}}|R|^{\frac{\nu}{2}}\Gamma_{n_d}\left(\frac{\nu}{2}\right)}, \tag{A1}$$

where ν is the degrees of freedom, and R the scale matrix. The mean is $E[S] = \nu R$, and the variance is $\mathrm{Var}[S_{ij}] = \nu[R_{ij}^2 - R_{ii}R_{jj}]$.

The inverse Wishart distribution

$$f_{W^{-1}}(R|S,\nu) = \frac{|S|^{\frac{\nu}{2}}|R|^{-\frac{\nu+n_d+1}{2}}\exp\left[-\frac{1}{2}Tr(R^{-1}S)\right]}{2^{\frac{\nu n_d}{2}}\Gamma_{n_d}\left(\frac{\nu}{2}\right)}, \tag{A2}$$

where ν is the degrees of freedom, and S the scale matrix. The mean is $E[R] = S/(\nu - n_d - 1)$.

The multivariate Normal or Gaussian distribution written in a form using the Trace operator

$$f_N(\boldsymbol{x}_0|\boldsymbol{\mu},R) = (2\pi)^{-\frac{n_d}{2}}|R|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}Tr\left(R^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu})(\boldsymbol{x}_0-\boldsymbol{\mu})^T\right)\right], \tag{A3}$$

with mean $E[\boldsymbol{x}_0] = \boldsymbol{\mu}$ and variance $\mathrm{Var}[\boldsymbol{x}_0] = R$.

The multivariate *t*-distribution

$$f_{t,\nu}(\boldsymbol{x}_0|\boldsymbol{\mu},R) = \frac{\Gamma[(\nu+n_d)/2]}{\Gamma(\nu/2)(\nu\pi)^{n_d/2}|R|^{1/2}} \times\left[1+(\boldsymbol{x}_0-\boldsymbol{\mu})^T(\nu R)^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu})\right]^{-\frac{\nu+n_d}{2}}, \tag{A4}$$

where ν is the degrees of freedom, and R the scale matrix. The mean is $E[\boldsymbol{x}_0] = \boldsymbol{\mu}$, and the variance is $\mathrm{Var}[\boldsymbol{x}_0] = \frac{\nu}{\nu-2}R$.

APPENDIX B: PERTURBATIVE BASED APPROACH FOR EXPRESSIONS INVOLVING THE COVARIANCE OF WISHART-DISTRIBUTED MATRICES

In this appendix, we consider the perturbation based approach to understanding the biases involved in a statistical analysis of data when the covariance matrix itself is a random variable S. To do this, we use the expressions between estimated and true covariance matrix as provided by Taylor, Joachimi & Kitching (2013). Let $(hS)^{-1} = \Sigma^{-1} + \Delta_{\Sigma^{-1}}$. As S is drawn from a Wishart distribution, the errors $\Delta_{\Sigma^{-1}}$ can be written

$$\langle(\Delta_{\Sigma^{-1}})_{ab}(\Delta_{\Sigma^{-1}})_{cd}\rangle_S = A\Sigma^{-1}_{ab}\Sigma^{-1}_{cd} + B\left(\Sigma^{-1}_{ac}\Sigma^{-1}_{bd} + \Sigma^{-1}_{ad}\Sigma^{-1}_{bc}\right), \tag{B1}$$

where

$$A = \frac{2}{(n_s-n_d-1)(n_s-n_d-4)},$$
$$B = \frac{(n_s-n_d-2)}{(n_s-n_d-1)(n_s-n_d-4)}. \tag{B2}$$

First, we consider a perturbative expansion of $F_S^{-1} = h^{-1}(F_\Sigma + \Delta_F)^{-1}$, with Δ_F defined as a standard Gaussian Fisher matrix with inverse covariance $\Delta_{\Sigma^{-1}}$ as required in Section 3.1. Expanding this, and taking the expected value, the first-order terms in Δ_F tend to zero (as $(hS)^{-1}$ is an unbiased estimator of Σ^{-1}), and so we are only interested in the second-order term in Δ_F, which can be written

$$\langle(F_\Sigma + \Delta_F)^{-1}\rangle_S\big|_{s.o.} = F_\Sigma^{-1}\Delta_F F_\Sigma^{-1}\Delta_F F_\Sigma^{-1}. \tag{B3}$$

Putting the relationships given in equation (B1) into equation (B3), we find that

$$\left\langle F_S^{-1}\right\rangle_S \simeq h^{-1}\left[1 + A + B(n_\theta+1)\right]F_\Sigma^{-1}. \tag{B4}$$

The calculation of the inverse Fisher matrix averaged over S using this perturbation based approach was performed in Percival et al. (2014) for the Gaussian Fisher matrix. As shown in the derivation leading to equation (32), this expression does not have to be solved perturbatively as an exact solution is possible. The non-perturbative solution is given in equation (31).

The next expression that we wish to understand perturbatively is $\langle F_S^{-1}Tr[S^{-1}\Sigma]\rangle_S$, as required in Section 3.2.2 and given in equation (45). The expression for which we are taking the expectation can be written

$$\left(F_S^{-1}Tr[S^{-1}\Sigma]\right)_{\alpha\beta} = \left[F_S^{-1}\right]_{\alpha\beta}S^{-1}_{ab}\Sigma_{ab}. \tag{B5}$$

The second-order term from F_S^{-1} is given by equation (B4), leading a term $n_d[1 + A + B(n_\theta + 1)]$, with the factor $n_d h$ coming from the summation over the term $S^{-1}_{ab}\Sigma_{ab}$. There is also a second-order cross term from F_S^{-1} and S^{-1}_{ab}, which gives $-[n_d A + 2B]$. Adding these together, we find the result in equation (45).

To approximate the expression in equation (46), note that there are eight possible ways that we can have pairs of $\Delta_{\Sigma^{-1}}$ in

$$\left[F_S^{-1}\right]_{\alpha\beta}\frac{\mathrm{d}\mu_a}{\mathrm{d}\theta_{\alpha'}}S^{-1}_{ab}\Sigma_{bc}S^{-1}_{cd}\frac{\mathrm{d}\mu_d}{\mathrm{d}\theta_{\beta'}}\left[F_S^{-1}\right]_{\beta'\alpha'}, \tag{B6}$$

with one at second order from each F_S^{-1}, the cross pair between the two F_S^{-1} and the cross pair from the two S^{-1}, and four cross pairs between F_S^{-1} and S^{-1}. Treating each in turn and expanding using equation (B1) leads to the result in equation (46).

Finally, we note that the expression in equation (50) can be derived similarly. To see this, note that there are eight possible ways that we

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

can have pairs of $\Delta_{\Sigma^{-1}}$ in

$$\langle\hat{\boldsymbol{\theta}}_\alpha\hat{\boldsymbol{\theta}}_\beta\rangle_{x,S} = \left[F_S^{-1}\right]_{\alpha\alpha'}\frac{\mathrm{d}\mu_a}{\mathrm{d}\theta_{\alpha'}}S^{-1}_{ab}\Sigma_{bc}S^{-1}_{cd}\frac{\mathrm{d}\mu_d}{\mathrm{d}\theta_{\beta'}}\left[F_S^{-1}\right]_{\beta'\beta}, \tag{B7}$$

similar to the expansion of equation (B6). These expressions are different - for example in the limit as $hS \rightarrow \Sigma$, equation (B6) tends towards $n_\theta(F_\Sigma^{-1})_{\alpha\beta}$, while equation (B7) tends towards $(F_\Sigma^{-1})_{\alpha\beta}$. Treating each of the eight possible combinations of two $\Delta_{\Sigma^{-1}}$ separately, expanding using equation (B1) and summing the terms gives the result in equation (50), which was the primary result of Dodelson & Schneider (2013).

APPENDIX C: COMPRESSING THE DATA

The effect of a linear compression of the data on model parameter inference can be considered using a property of the multivariate *t*-distribution. For some $n_c \times n_d$ matrix M, assuming that

$$f(\boldsymbol{\mu}|\boldsymbol{x}_0, S) = f_{t,m-n_d}\left(\boldsymbol{\mu}\,\middle|\,\boldsymbol{x}_0, \frac{n_s-1}{m-n_d}S\right), \tag{C1}$$

then a property of the multivariate t-distribution is that

$$f(M\boldsymbol{\mu}|M\boldsymbol{x}_0, S) = f_{t,m-n_d}\left(M\boldsymbol{\mu}\,\middle|\,M\boldsymbol{x}_0, \frac{n_s-1}{m-n_d}MSM^T\right). \tag{C2}$$

Now consider an analysis of the compressed data, where we apply a compression with $n_c = n_\theta$ and

$$M = F_S^{-1}\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\theta}}^T S^{-1}, \tag{C3}$$

such that the MAP estimate $\hat{\boldsymbol{\theta}} = M\boldsymbol{x}_0$, and

$$MSM^T = F_S^{-1}\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\theta}}^T S^{-1}SS^{-1}\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\theta}}F_S^{-1} = F_S^{-1}. \tag{C4}$$

For data analysed with a Gaussian posterior and linear model, such a compression is sufficient in that the analysis of the reduced data gives the same inferences as those from the full data set, including the covariance on $\boldsymbol{\mu}$. Assuming a t-distribution posterior for $\boldsymbol{x}_0$, we find that the posterior for the reduced data is

$$f(M\boldsymbol{\mu}|M\boldsymbol{x}_0, S) = f_{t,m-n_d}\left(M\boldsymbol{\mu}\,\middle|\,M\boldsymbol{x}_0, \frac{n_s-1}{m-n_d}MSM^T\right). \tag{C5}$$

Now, defining $\boldsymbol{\theta}' = M\boldsymbol{\mu}$, as an estimator for the MAP values, we see that

$$f(\boldsymbol{\theta}'|\hat{\boldsymbol{\theta}}, S) = f_{t,m-n_d}\left(\boldsymbol{\theta}'\,\middle|\,\hat{\boldsymbol{\theta}}, \frac{n_s-1}{m-n_d}F_S^{-1}\right). \tag{C6}$$

This gives that the covariance for $\boldsymbol{\theta}'$ is

$$\langle(\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}})^T\rangle = \frac{n_s-1}{m-n_d-2}F_S^{-1}. \tag{C7}$$

This is the covariance recovered from the compressed data as given by equation (C3), for a measurement of the MAP estimates $\hat{\boldsymbol{\theta}}$. This does not match the expression in equation (38), but does match the solution of equation (41) where we integrate under the posterior and then average over $\boldsymbol{x}_0$, assuming that this was drawn from a multivariate t-distribution.

Our interpretation of this is that the linear compression of the data analysed with a *t*-distribution posterior does not include information about the distribution of the data around the MAP estimate, as is used in equation (38) to determine a specific model parameter covariance for that realization of the data. Without this extra information, compressing the data means that the model parameter covariance recovered corresponds to the average for a distribution of $\boldsymbol{x}_0$, rather than that for a particular $\boldsymbol{x}_0$ recovered if using more data. Furthermore, the model parameter covariance corresponds to that recovered on average for data distributed according to a multivariate t-distribution. We therefore conclude that data compression works differently than when analysing using a Gaussian posterior for which linear compression is sufficient in terms of giving the same MAP estimate and covariance. For multivariate t-distribution posteriors, this is not the case, and additional information is used on the distribution of the data around the MAP estimate in order to determine the model parameter covariance as shown in equation (38). This will be considered further in future work.

APPENDIX D: INTERPRETATION OF CREDIBLE INTERVALS BASED ON σ

We now consider how the use of a multivariate t-distribution affects the interpretation of confidence intervals. Where credible intervals are derived directly from the posterior, for example, by considering the fraction of points within a given interval for an MCMC chain exploring a posterior volume, then the interpretation of results is correct whatever the form of the posterior. However, if one wants to define or interpret intervals based on ±$N\sigma$ contours, then one needs to be careful when interpreting a posterior with t-distribution form, as explored in this Appendix.

As our favoured solution assumes a power-law prior, the posterior, when written in terms of the model parameters for linear models, has a multivariate t-distribution form with degrees of freedom $\nu = m - n_\theta$. When marginalized over other parameters, the posterior probability for each model parameter has a form matching the student t-distribution for the parameter $\sqrt{(\nu-2)/\nu}\boldsymbol{\theta}_i/\sigma$. In general, the t-distribution has broader tails and a narrower core than the Gaussian distribution, matching the Gaussian distribution in the limit $\nu \rightarrow \infty$. The variance of the standard t-distribution is $\nu/(\nu-2)$, and so we need a broader range of integration to determine a ±$N\sigma$ interval, integrating over $\pm N\sqrt{\nu/(\nu-2)}$ rather than ±N as with a Gaussian for distribution with unit variance. The probabilities associated with credible intervals based on ±$N\sigma$ are compared in Fig. D1: the ±1σ credible interval is more probable for the *t*-distribution compared with the Gaussian distribution with the same variance. However, the tail probabilities are larger for the *t*-distribution than the Gaussian to fixed ±$N\sigma$ limits for $N \geq 2$. Fig. D2 instead shows the change in N

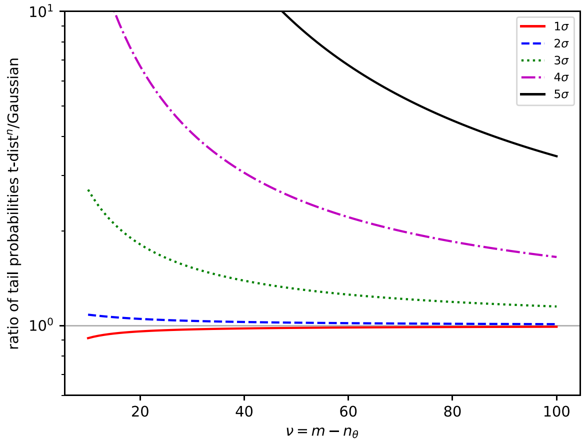

Figure D1. The ratio of tail probabilities for t-distribution and Gaussian posteriors outside of ±$N\sigma$ credible intervals, where $N = 1\ldots\ 5$.

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

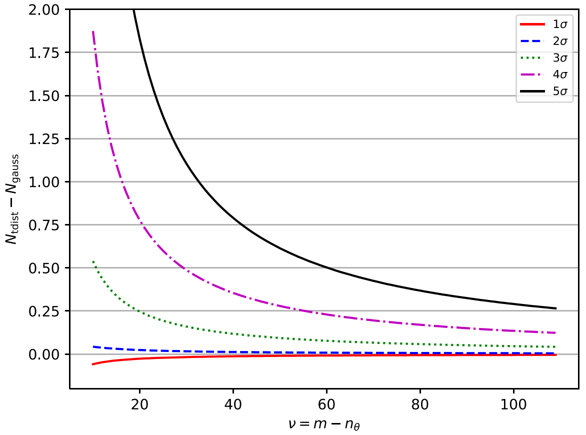

Figure D2. The difference between the linear factors required to define credible intervals containing a fixed probability for t-distribution and Gaussian posteriors. Intervals are defined using the Gaussian probability within the $\pm N\sigma$ interval, where $N = 1... 5$. So, for example, tracing the 5σ curve (solid black line), we see that for $\nu = 100$, to match the Gaussian $\pm 5\sigma$ coverage probability, we would need to consider a $\pm 5.29\sigma$ interval for the *t*-distribution.

required to match tail probabilities from the t-distribution to those from the Gaussian distribution. For example, with a t-distribution posterior with $\nu = 100$, one would need to define an interval based on the $\pm 5.29\sigma$ threshold to match the inference (including tail probabilities) made from a 5σ result with a Gaussian posterior. We would therefore need to integrate to larger intervals for the t-distribution to reduce the tail probabilities to match the Gaussian values for $N \geq 2$. For smaller ν we need to integrate to larger intervals in σ.

APPENDIX E: ANALYTIC MARGINALIZATION FOR ESTIMATING THE MEAN OF DATA

In this appendix we outline the derivations that allow us to significantly speed up our Monte Carlo simulations fitting a single mean value $\bar{\mu}$ to n_d correlated data values $\boldsymbol{x}_0$, and ultimately would make them superfluous as we could perform all of the necessary calculations analytically. These are a special case of the derivation given in Section 3.2.2 and are therefore not strictly necessary, but we include it as we feel that it gives insight into the problem being solved. To help with this, we first consider the more familiar case of a Gaussian posterior.

E1 Fitting the mean with a multivariate Gaussian posterior

We start with the simple case of a Gaussian posterior. For this, we can use the standard definition of $\chi^2 = -2\ln L$ for fitting a mean $\bar{\mu}$ to data $\boldsymbol{x}_0$ with inverse covariance matrix $(hS)^{-1}$

$$\chi^2 \equiv \sum_{ij} ((\boldsymbol{x}_0)_i - \bar{\mu})(hS)^{-1}_{ij}((\boldsymbol{x}_0)_j - \bar{\mu}). \tag{E1}$$

Expanding, we can write

$$\chi^2 = h^{-1}[C_1 - 2C_2\bar{\mu} + C_3\bar{\mu}^2], \tag{E2}$$

where

$$C_1 = \sum_{ij} (x_0)_i S^{-1}_{ij} (x_0)_j, \tag{E3}$$

$$C_2 = \sum_{ij} S^{-1}_{ij} (x_0)_j, \tag{E4}$$

$$C_3 = \sum_{ij} S^{-1}_{ij}. \tag{E5}$$

To align with the notation used elsewhere in this paper, we note that for this problem, the parameter $\theta = \bar{\mu}$, the model is $\mu_i = \bar{\mu}$, and we have $d\mu_i/d\theta = 1$, $F_S = C_3$ and $F_S^{-1} = 1/C_3$. The derivative $\mathrm{d}\boldsymbol{\mu}/\mathrm{d}\boldsymbol{\theta} = U$, where the unit vector U is a vector of 1's, and $C_2 = U^T S^{-1}\boldsymbol{x}_0$. We now 'complete the square' for the model-dependent part of χ^2

$$-2C_2\bar{\mu} + C_3\bar{\mu}^2 = C_3\left(\bar{\mu} - \frac{C_2}{C_3}\right)^2 - \frac{C_2^2}{C_3}. \tag{E6}$$

We can then write the posterior as a Gaussian distribution around the MAP estimate

$$f(\bar{\mu}|\boldsymbol{x}_0, S) \propto \exp\left[-\frac{C_3}{2h}\left(\bar{\mu} - \frac{C_2}{C_3}\right)^2\right]. \tag{E7}$$

The mean, as derived from the posterior therefore has a Gaussian distribution, and the expected value for $\bar{\mu}$ and the variance can then be read off, $\langle\bar{\mu}\rangle_x = C_2/C_3$, and $\sigma^2 = h/C_3$. As expected for a Gaussian posterior and a linear model, the MAP estimate matches the value given in equation (48), and the expected model parameter variance integrating under the posterior matches the inverse of the Fisher matrix. So we see that a Gaussian fit to the peak of the posterior also describes the results from the full distribution.

E2 Fitting the mean with multivariate *t*-distribution posterior

This section replicates Section 3.2.2, but now for the special case of fitting the mean to a set of data, as considered in Section 5. We do this as we used these equations to speed-up the Monte Carlo runs presented in Section 5, and in order to allow them to be used as an aide to understanding the derivation in Section 3.2.2. Consequently, we try to keep the layout and structure similar and make no apologies for replication. We only present the derivation for Gaussian distributed data.

Assuming that the posterior has a scale matrix $(n_s - 1)/(m - n_d)S$, and degrees of freedom $\nu = m - n_d$, as in equation (22) we can write the posterior where the model is a constant mean value $\bar{\mu}$

$$f(\bar{\mu}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{1}{n_s - 1}\sum_{ij}((\boldsymbol{x}_0)_i - \bar{\mu})\, S^{-1}_{ij}((\boldsymbol{x}_0)_j - \bar{\mu})\right]^{-\frac{m}{2}}. \tag{E8}$$

Expanding as in the Gaussian case, we have

$$f(\bar{\mu}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{1}{n_s - 1}(C_1 - 2C_2\bar{\mu} + C_3\bar{\mu}^2)\right]^{-\frac{m}{2}}, \tag{E9}$$

and completing the square gives

$$f(\bar{\mu}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{1}{n_s - 1}\left(C_1 - \frac{C_2^2}{C_3}\right) + \frac{C_3}{n_s - 1}\left(\bar{\mu} - \frac{C_2}{C_3}\right)^2\right]^{-\frac{m}{2}}. \tag{E10}$$

We now define

$$y = \sqrt{C_3}\left(\bar{\mu} - \frac{C_2}{C_3}\right)\left[1 + \frac{1}{n_s - 1}\left(C_1 - \frac{C_2^2}{C_3}\right)\right]^{-1/2} \times \left(\frac{n_s - 1}{m - 1}\right)^{-1/2}, \tag{E11}$$

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022

so that

$$f(\bar{\mu}|\boldsymbol{x}_0, S) \propto \left[1 + \frac{y^2}{m-1}\right]^{-\frac{m}{2}}. \tag{E12}$$

We see that y is distributed with a t-distribution with $m - 1$ degrees of freedom, such that the mean $\langle y\rangle = 0$, and the variance $\langle y^2\rangle = (m - 1)/(m - 3)$.

We can write $\bar{\mu}$ in the form $\bar{\mu} = ay + b$, which has the property that $\langle\bar{\mu}\rangle = a\langle y\rangle + b$, and $\mathrm{Var}(\bar{\mu}) = a^2\mathrm{Var}(y)$:

$$\bar{\mu} = \frac{1}{\sqrt{C_3}}\left[1 + \frac{1}{n_s - 1}\left(C_1 - \frac{C_2^2}{C_3}\right)\right]^{1/2}\left(\frac{n_s - 1}{m - 1}\right)^{1/2} y + \frac{C_2}{C_3}. \tag{E13}$$

From this, we see that the distribution of $\bar{\mu}$ has mean $\langle\bar{\mu}\rangle = C_2/C_3$, as expected given the discussion in Section 3.3. The variance for any realization is

$$\langle(\bar{\mu} - \langle\bar{\mu}\rangle)^2\rangle = \frac{n_s - 1}{m - 3}\frac{1}{C_3}\left[1 + \frac{1}{n_s - 1}\left(C_1 - \frac{C_2^2}{C_3}\right)\right], \tag{E14}$$

which matches equation (38) for a fit to the mean. Thus, rather than numerically integrate under the posterior for any realization of $\boldsymbol{x}_0$ and S, we can instead use this expression for the variance recovered. We have confirmed numerically that this result is correct, and that the variance depends on the data as given in this equation. Unlike for the Gaussian distribution, here the recovered variance depends on the value of $\boldsymbol{x}_0$ through C_1 and C_2. These terms do not cancel in general.

We can now consider the expected value, averaging over multiple sets of data, but using the same covariance matrix approximation S to determine the posterior. In this case, C_3 is fixed, and we need to replace the terms C_1 and C_2^2 by the relevant expected values. Remembering that $\boldsymbol{x}$ was drawn from a Gaussian distribution with covariance Σ and zero mean, we have

$$\langle C_1\rangle_x = \sum_{ij}[S^{-1}\Sigma]_{ij}, \tag{E15}$$

$$\left\langle C_2^2\right\rangle_x = \sum_{ij}[S^{-1}\Sigma S^{-1}]_{ij}. \tag{E16}$$

This is the expected result for the variance recovered for many Gaussian distributed realizations of the data $\boldsymbol{x}_0$. equation (E14), together with the expressions of equations (E15) & E16, allow us not to run Monte Carlo simulations for different data for the same covariance, as we can accurately predict the result using these equations.

To go one step further when finding analytic expressions for the Monte Carlo runs, we now need to find expressions for the relevant terms in equation (E14), now considering the expected values averaging over all possible covariance matrices S. We can do this using the expressions given in Section B for the expansion of S around the true matrix Σ. These give

$$\langle 1/C_3\rangle_S = [1 + A + 2B]h^{-1}F_\Sigma^{-1}, \tag{E17}$$

$$\langle C_1/C_3\rangle_S = [n_d + 2B(n_d - 1)]F_\Sigma^{-1}, \tag{E18}$$

$$\left\langle C_2^2/C_3^2\right\rangle_S = [1 + B(n_d - 1)]F_\Sigma^{-1}. \tag{E19}$$

As expected, this final two equations match equations (45) & 46 with $n_\theta = 1$. For the first expression we write here the perturbative result rather than the exact form as used in Section 3.2.1. In terms of n_θ, this is $\langle F_S^{-1}\rangle = h^{-1}[1 + A + B(n_\theta + 1)]F_\Sigma^{-1}$. Note that by using these expressions we would have removed any need to do the Monte Carlo simulations, as we have analytic expressions for all stages of the Monte Carlo runs being performed, albeit to second order in the covariance matrix approximation.

We can also consider how, for this case of fitting the mean to correlated data, we can derive an analytic expression for the scatter in recovered MAP estimates. To determine this, note from equation (E9) that the MAP estimate (obtained by taking the log and setting the derivative with respect to $\bar{\mu}$ to zero in the posterior) is C_2/C_3. From the definition of these quantities, $C_2/C_3 = U^T S^{-1}\boldsymbol{x}_0/C_3$. Remembering that $\boldsymbol{x}_0$ are drawn from a Gaussian distribution with covariance Σ, we see that C_2/C_3 is also Gaussian distributed with zero mean and variance $U^T S^{-1}\Sigma S^{-1}U/C_3^2 = \langle C_2^2/C_3^2\rangle_S$. Equation (50) then shows that this matches the Dodelson & Schneider (2013) result.

A reader having reached this stage of the paper firstly needs congratulating, but also might well be asking why we need to run the Monte Carlo simulations presented in Section 5 at all given that we have analytically approximated all of the results we will extract from those simulations. And they would be correct. However, we keep Fig. 2 as it adds colour and confirms the validity of the approximations – using the Fisher matrix to determine confidence intervals from the posterior, and the second-order expansions through which we estimated the impact of S.

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

Downloaded from https://academic.oup.com/mnras/article/510/3/3207/6449384 by Imperial College London Library user on 24 March 2022