Article

An Analytical Approach to Bayesian Evidence Computation

Juan García-Bellido (ORCID)

Departamento de Física Teórica C-XI, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain; juan.garciabellido@uam.es

Abstract: Bayesian evidence is a key tool in model selection, allowing a comparison of models with different numbers of parameters. Its use in the analysis of cosmological models has been limited by difficulties in calculating it, with current numerical algorithms requiring supercomputers. In this paper we give exact formulae for the Bayesian evidence in the case of Gaussian likelihoods with arbitrary correlations and top-hat priors, and approximate formulae for the case of likelihood distributions with leading non-Gaussianities (skewness and kurtosis). We apply these formulae to cosmological models with and without isocurvature components, and compare with results we previously obtained using numerical thermodynamic integration. We find that the results are of lower precision than the thermodynamic integration, while still being good enough to be useful.

Keywords: Bayesian evidence; Gaussian likelihoods; thermodynamic integration

1. Introduction

Model selection refers to the statistical problem of deciding which model description of observational data is the best [1,2]. It differs from parameter estimation, where the choice of a single model (i.e., choice of parameters to be varied) has already been made and the aim is to find their best-fitting values and ranges. While there have been widespread applications of parameter estimation techniques, usually likelihood fitting, to cosmological data, there has so far been quite limited application of model selection statistics [3–12]. This is unfortunate, as model selection techniques are necessary to robustly distinguish between models with different numbers of parameters, and many of the most interesting issues in cosmology concern the desirability or otherwise of incorporating additional parameters to describe new physical effects.

Within the context of Bayesian inference, model selection should be carried out using the *Bayesian evidence* [1,2], which measures the probability of the model in light of the observational data (i.e., the average likelihood over the prior distribution). The Bayesian evidence associates a single number with each model, and the models can then be ranked in order of the evidence, with the ratios of those values interpreted as the relative probability of the models. This process sets up a desirable tension between model simplicity and the ability to fit the data.

Use of the Bayesian evidence has so far been limited by difficulties in calculating it. The standard technique is thermodynamic integration [13,14], which varies the temperature in a Monte Carlo Markov Chain (MCMC) approach in order that the distribution is sampled in a way covering both posterior and prior distributions. However, in recent work [12] we showed that in order to obtain sufficiently-accurate results in a cosmological context, around 10^7 likelihood evaluations are required per model. Such analyses are CPU-limited by the time needed to generate the predicted spectra to compare with the data, and this requirement pushes the problem into the supercomputer class (for comparison, parameter estimation runs typically employ 10^5 to 10^6 likelihood evaluations).

In this paper, we propose and exploit a new analytic method to compute the evidence based on an expansion of the likelihood distribution function. The method pre-supposes that the covariance of the posterior distribution has been obtained, for instance via an

MCMC parameter estimation run, and in its present form requires that the prior distributions of the parameters are uniform top-hat priors.[1] While the method will not be applicable for general likelihood distributions, we include the leading non-Gaussianities (skewness and kurtosis) in approximating the likelihood shape, with the expectation of obtaining good results whenever the likelihood distribution is sufficiently simple. Cosmological examples commonly exhibit likelihood distributions with only a single significant peak.

We apply the method both to toy model examples and to genuine cosmological situations. In particular, we calculate the evidences for adiabatic and isocurvature models, which we previously computed using thermodynamic integration in ref. [12]. We find that the discrepancies between the methods are typically no worse than 1 in ln(Evidence), meaning that the analytical method is somewhat less accurate than would be ideal, but is accurate enough to give a useful indication of model preference.

2. The Bayesian Evidence

The posterior probability distribution $\mathcal{P}(\theta, \mathcal{M}|\mathbf{D})$ for the parameters θ of the model $\mathcal{M}$, given the data $\mathbf{D}$, is related to the likelihood function $\mathcal{L}(\mathbf{D}|\theta, \mathcal{M})$ within a given set of prior distribution functions $\pi(\theta, \mathcal{M})$ for the parameters of the model, by Bayes' theorem:

$$\mathcal{P}(\theta, \mathcal{M}|\mathbf{D}) = \frac{\mathcal{L}(\mathbf{D}|\theta, \mathcal{M})\, \pi(\theta, \mathcal{M})}{E(\mathbf{D}|\mathcal{M})}, \tag{1}$$

where E is the Bayesian evidence, i.e., the average likelihood over the priors,

$$E(\mathbf{D}|\mathcal{M}) = \int d\theta\, \mathcal{L}(\mathbf{D}|\theta, \mathcal{M})\, \pi(\theta, \mathcal{M}), \tag{2}$$

where θ is a vector with n-components characterising the n independent parameters. The prior distribution function π contains all the information about the parameters before observing the data, i.e., our theoretical prejudices, our physical understanding of the model, and input from previous experiments.

In the case of a large number of parameters ($n \gg 1$), the evidence integral cannot be performed straightforwardly and must be obtained either numerically or via an analytical approximation. Amongst numerical methods the most popular is thermodynamic integration [13,14] but this can be computationally intensive [12]. An alternative is the application of the nested sampling algorithm [15,16] and Monte Carlo methods with the stepping-stone sampling algorithm [17,18]. On the other hand, the simplest analytical approximation is the Laplace approximation, valid when the distribution can be approximated by a multivariate Gaussian. This may hold when the quantity and quality of the data is optimal, but is likely to be valid only in limited cosmological circumstances.

The Bayesian evidence is of interest because it allows a comparison of models amongst an exclusive and exhaustive set $\{\mathcal{M}_i\}_{i=1...N}$. We can compute the posterior probability for each hypothesis given the data $\mathbf{D}$ using Bayes' theorem:

$$\mathcal{P}(\mathcal{M}_i|\mathbf{D}) \propto E(\mathbf{D}|\mathcal{M}_i)\, \pi(\mathcal{M}_i), \tag{3}$$

where $E(\mathbf{D}|\mathcal{M}_i)$ is the evidence of the data under the model $\mathcal{M}_i$, and $\pi(\mathcal{M}_i)$ is the prior probability of the ith model before we see the data. The ratio of the evidences for the two competing models is called the Bayes factor [19]

$$B_{ij} = \frac{E(\mathbf{D}|\mathcal{M}_i)}{E(\mathbf{D}|\mathcal{M}_j)}, \tag{4}$$

and this is also equal to the ratio of the posterior model probabilities if we assume that we do not favour any model a priori, so that $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2) = \ldots = \pi(\mathcal{M}_N) = 1/N$.

The Bayes factor Equation (4) provides a mathematical representation of Occam's razor, because more complex models tend to be less predictive, lowering their average

likelihood in comparison to simpler, more predictive models. More complex models can only be favoured if they are able to provide a significantly improved fit to the data. In simple cases where models give vastly different maximum likelihoods there is no need to employ model selection techniques, but they are essential for properly discussing cases where the improvement of fit is marginal. This latter situation is more or less inevitable whenever the possibility of requiring an additional parameter arises from new data, unless the new data is of vastly greater power than that preceding it; cosmological examples include the inclusion of spectral tilt, dark energy density variation, or trace isocurvature perturbations, explored later in this paper.

In this paper we will obtain analytical formulae which approximates the Bayesian evidence by considering the higher-order cumulants of the distribution in a systematic way. The advantage is that with these analytical formulae one can compute the evidence for a given model with an arbitrary number of parameters, given the hierarchy of cumulants of the distribution, assumed previously computed for the likelihood distribution function within the parameter estimation programme.

The evidence needs to be calculated to sufficient precision for robust conclusions to be drawn. The standard interpretational scale, due to Jeffreys [1] and summarized in ref. [12], strengthens its verdict roughly each time the difference in ln(Evidence) increases by one. The evidence therefore needs to be computed more accurately than this, with an uncertainty of 0.1 in ln(Evidence) easily sufficient, and a factor two worse than that acceptable. This accuracy requirement ensures that the relative model probabilities are changed little by the uncertainty.

The first thing needed is to characterize the distribution function for the model with n parameters. Let $f(\mathbf{x})$ be this function, and let us assume that it is properly normalized,

$$\int_{-\infty}^{\infty} d^n\mathbf{x}\, f(\mathbf{x}) = 1\,. \tag{5}$$

Then, the p-point correlation function is given by

$$\langle x_{i_1} \ldots x_{i_p} \rangle = \int_{-\infty}^{\infty} d^n\mathbf{x}\, x_{i_1} \ldots x_{i_p}\, f(\mathbf{x})\,. \tag{6}$$

From this distribution function one can always construct the generating functional, $\phi(\mathbf{u})$, as the Fourier transform

$$\phi(\mathbf{u}) = \int_{-\infty}^{\infty} d^n\mathbf{x}\, e^{i\,\mathbf{u}\cdot\mathbf{x}}\, f(\mathbf{x})\,. \tag{7}$$

This function can be expanded as

$$\phi(\mathbf{u}) = \exp\left[\sum_{p=1}^{\infty} \frac{i^p}{p!}\, A_{i_1\ldots i_p}\, u^{i_1} \ldots u^{i_p}\right]\,, \tag{8}$$

where $A_{i_1\ldots i_p}$ are totally symmetric rank-p tensors. For instance, if we restrict ourselves to order 4, we can write

$$\phi(\mathbf{u}) = \exp\left[i\,\mu_i u_i - \frac{1}{2!}\, C_{ij}\, u_i u_j - \frac{i}{3!}\, B_{ijk}\, u_i u_j u_k + \frac{1}{4!}\, D_{ijkl}\, u_i u_j u_k u_l + \cdots + \frac{i^n}{n!}\, A_{i_1\ldots i_n}\, u_{i_1} \ldots u_{i_n}\right]\,, \tag{9}$$

where μ_i is the mean value of variable x_i; C_{ij} is the covariance matrix; B_{ijk} is the trilinear matrix associated with the third cumulant or skewness; D_{ijkl} is the rank-4 tensor associated with the fourth cumulant or kurtosis; and $A_{i_1\ldots i_n}$ is the rank-n tensor associated with the n-th cumulant. Their expressions in terms of n-point correlation functions can be obtained from Equation (7), by realising that

$$\langle x_{i_1} \ldots x_{i_n} \rangle = (-i)^n\, \frac{\partial^n \phi(\mathbf{u})}{\partial u_{i_1} \ldots \partial u_{i_n}}\bigg|_{\mathbf{u}=0}\,. \tag{10}$$

For instance, the first-order term gives

$$\langle x_i \rangle = (-i) \left. \frac{\partial \phi(\mathbf{u})}{\partial u_i} \right|_{\mathbf{u}=0} = \mu_i \,. \tag{11}$$

The second-order correlation function gives

$$\langle x_i x_j \rangle = (-i)^2 \left. \frac{\partial^2 \phi(\mathbf{u})}{\partial u_i \partial u_j} \right|_{\mathbf{u}=0} = C_{ij} + \mu_i \mu_j \,, \tag{12}$$

such that the covariance matrix is obtained, as usual, from

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \,.$$

The third-order correlation function gives

$$\langle x_i x_j x_k \rangle = (-i)^3 \left. \frac{\partial^3 \phi(\mathbf{u})}{\partial u_i \partial u_j \partial u_k} \right|_{\mathbf{u}=0} = B_{ijk} + \mu_i C_{jk} + \mu_j C_{ki} + \mu_k C_{ij} + \mu_i \mu_j \mu_k \,, \tag{13}$$

such that the skewness matrix is obtained from

$$B_{ijk} = \langle x_i x_j x_k \rangle - \langle x_i \rangle \langle x_j x_k \rangle - \langle x_j \rangle \langle x_k x_i \rangle - \langle x_k \rangle \langle x_i x_j \rangle + 2 \langle x_i \rangle \langle x_j \rangle \langle x_k \rangle \,. \tag{14}$$

The fourth-order correlation function gives

$$
\begin{aligned}
\langle x_i x_j x_k x_l \rangle = (-i)^4 \left. \frac{\partial^4 \phi(\mathbf{u})}{\partial u_i \partial u_j \partial u_k \partial u_l} \right|_{\mathbf{u}=0} =\ & D_{ijkl} + C_{ij} C_{kl} + C_{ik} C_{jl} + C_{il} C_{jk} \\
+\ & B_{ijk} \mu_l + B_{ijl} \mu_k + B_{jkl} \mu_i + B_{ikl} \mu_j \\
+\ & C_{ij} \mu_k \mu_l + C_{ik} \mu_j \mu_l + C_{il} \mu_j \mu_k \\
+\ & C_{jk} \mu_i \mu_l + C_{jl} \mu_i \mu_k + C_{kl} \mu_i \mu_j \\
+\ & \mu_i \mu_j \mu_k \mu_l \,,
\end{aligned}
\tag{15}
$$

such that the kurtosis matrix is obtained from

$$
\begin{aligned}
D_{ijkl} =\ & \langle x_i x_j x_k x_l \rangle - \langle x_i x_j \rangle \langle x_k x_l \rangle - \langle x_i x_k \rangle \langle x_j x_l \rangle - \langle x_i x_l \rangle \langle x_j x_k \rangle \\
-\ & \langle x_i x_j x_k \rangle \langle x_l \rangle - \langle x_i x_j x_l \rangle \langle x_k \rangle - \langle x_i x_k x_l \rangle \langle x_j \rangle - \langle x_j x_k x_l \rangle \langle x_i \rangle \\
+\ & 2 \langle x_i x_j \rangle \langle x_k \rangle \langle x_l \rangle + 2 \langle x_i x_k \rangle \langle x_j \rangle \langle x_l \rangle + 2 \langle x_i x_l \rangle \langle x_j \rangle \langle x_k \rangle + 2 \langle x_j x_k \rangle \langle x_i \rangle \langle x_l \rangle \\
+\ & 2 \langle x_j x_l \rangle \langle x_i \rangle \langle x_k \rangle + 2 \langle x_k x_l \rangle \langle x_i \rangle \langle x_j \rangle - 6 \langle x_i \rangle \langle x_j \rangle \langle x_k \rangle \langle x_l \rangle \,,
\end{aligned}
\tag{16}
$$

and so on, for the higher-order cumulants.

3. The Gaussian Approximation

Let us first evaluate the evidence for a multivariate Gaussian distribution, that is, one in which all the cumulants are zero except the covariance matrix C_{ij} and the means μ_i. In this case, the generating functional and the distribution are given by [2]

$$\phi(\mathbf{u}) = \exp\left[-i\mu_i u_i - \frac{1}{2} C_{ij} u_i u_j \right], \tag{17}$$

$$f(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} d^n\mathbf{u} \, e^{-i\mathbf{u}\cdot\mathbf{x}} \, \phi(\mathbf{u}) \tag{18}$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\left[-\frac{1}{2} C_{ij}^{-1}(x_i - \mu_i)(x_j - \mu_j) \right], \tag{19}$$

which satisfies

$$\langle x_i \rangle = \mu_i, \qquad \langle x_i x_j \rangle = C_{ij} + \mu_i \mu_j, \qquad \langle x_i x_j x_k \rangle = \mu_{(i} C_{jk)} + \mu_i \mu_j \mu_k, \qquad \dots \tag{20}$$

where the sub-indices in parenthesis, (ijk), indicate a cyclic sum. Notice that all the n-point correlation functions can be written in terms of the first two moments of the distribution, and all the higher-order cumulants vanish.

3.1. Centred Priors

For initial calculations, we assume a top-hat prior and make the unrealistic assumption, to be lifted later, that it is centred at the mean value:

$$\pi(x, a) \equiv \begin{cases} (2a)^{-1} & -a < x - \mu < a, \\ 0 & \text{otherwise}. \end{cases} \tag{21}$$

Since the Fourier transform of a top-hat function is

$$\int_{-\infty}^{\infty} dx \, e^{iux} \, \pi(x, a) = \frac{\sin au}{au} \exp[i\mu u],$$

we can write the evidence either way

$$E(a_1, \dots, a_n) = \int_{-\infty}^{\infty} d^n\mathbf{x} \, f(\mathbf{x}) \prod_{i=1}^{n} \pi(x_i, a_i) = \prod_{i=1}^{n} (2a_i)^{-1} \int_{-a_1}^{a_1} dx_1 \cdots \int_{-a_n}^{a_n} dx_n \, f(\tilde{\mathbf{x}}) \tag{22}$$

$$= \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} d^n\mathbf{u} \, \phi(\mathbf{u}) \prod_{i=1}^{n} \frac{\sin a_i u_i}{a_i u_i}. \tag{23}$$

In Equation (22) we integrate over the displaced coordinate, $\tilde{x}_i \equiv x_i - \mu_i$, such that $\langle \tilde{x}_i \rangle = 0$ and $\langle \tilde{x}_i \tilde{x}_j \rangle = C_{ij}$. From now on, we ignore the tildes, and assume we have moved to those coordinates. Note that the choice of prior is not crucial. We could have chosen a Gaussian prior, and the result would not be very different, except that the window functions, $\sin z / z$, would then be Gaussian. Let us now perform the integration Equation (22) in the case of one, two and then n variables.

One variable. Suppose the covariance is just $C = \sigma^2$. The evidence is then

$$E(a) = \frac{1}{2a\,\sigma\sqrt{2\pi}} \int_{-a}^{a} dx \, e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} du \, \frac{\sin au}{au} e^{-\frac{1}{2}\sigma^2 u^2} = \frac{1}{2a}\text{Erf}\left[\frac{a}{\sigma\sqrt{2}}\right], \tag{24}$$

where $\text{Erf}[x]$ is the error function, which asymptotes very quickly to one for $x \geq 2$, or $a \geq 3\sigma$. Therefore, the evidence of a model with centred top-hat prior of width $2a$ is well approximated by $(2a)^{-1}$. Note that the Bayesian evidence depends very strongly on the prior chosen for the model, and often choosing this prior is crucial for model specification [20].

Two variables. Suppose we have two correlated variables, x_1 and x_2, with covariance matrix

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{25}$$

where the cross-correlation ρ is defined by

$$\rho = \frac{\langle x_1 x_2 \rangle}{\sqrt{\langle x_1^2 \rangle \langle x_2^2 \rangle}} = \frac{\langle x_1 x_2 \rangle}{\sigma_1 \sigma_2},$$

with σ_1 and σ_2 as the corresponding quadratic dispersions. In this case, the normalized two-dimensional distribution function is

$$f(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[\frac{-1}{1-\rho^2}\left(\frac{x_1^2}{2\sigma_1^2} - \frac{\rho x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{2\sigma_2^2}\right)\right], \tag{26}$$

which has the property that integrating ("marginalizing") over one of the two variables, leaving a properly-normalized Gaussian distribution for the remaining variable,

$$\int_{-\infty}^{\infty} dx_2 \, f(\mathbf{x}) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{x_1^2}{2\sigma_1^2}}. \tag{27}$$

Let us now evaluate the evidence Equation (22) by integrating first over the prior in x_2,

$$\frac{1}{2a_2}\int_{-a_2}^{a_2} dx_2 \, f(\mathbf{x}) = \frac{e^{-\frac{x_1^2}{2\sigma_1^2}}}{\sigma_1\sqrt{2\pi}} \cdot \frac{1}{4a_2}\left[\mathrm{Erf}\left[\frac{a_2\sigma_1 + \rho\sigma_2\,x_1}{\sigma_1\sigma_2\sqrt{2(1-\rho^2)}}\right] + \mathrm{Erf}\left[\frac{a_2\sigma_1 - \rho\sigma_2\,x_1}{\sigma_1\sigma_2\sqrt{2(1-\rho^2)}}\right]\right]. \tag{28}$$

The first term is the result we would have obtained if we had been marginalizing over x_2; the second is a sum of error functions that still depend on x_1, and modulates the marginalization. We can use the series expansion of the error function to second order,

$$\frac{1}{2}\left(\mathrm{Erf}[a+x] + \mathrm{Erf}[a-x]\right) = \mathrm{Erf}[a] - \frac{2a\,x^2}{\sqrt{\pi}} e^{-a^2} + \mathcal{O}(x^4),$$

to write Equation (28) to order x_1^2 as

$$\frac{1}{2a_2}\int_{-a_2}^{a_2} dx_2 \, f(\mathbf{x}) = \frac{e^{-\frac{x_1^2}{2\sigma_1^2}}}{\sigma_1\sqrt{2\pi}}\left[\frac{1}{2a_2}\mathrm{Erf}\left[\frac{a_2}{\sigma_2\sqrt{2(1-\rho^2)}}\right] - \frac{\rho^2 x_1^2 e^{-\frac{a_2^2}{2\sigma_2^2(1-\rho^2)}}}{2\sigma_1^2\sigma_2(1-\rho^2)\sqrt{2\pi(1-\rho^2)}}\right]. \tag{29}$$

Integrating now over the x_1 prior, we finally obtain the evidence

$$\begin{aligned}
E(a_1, a_2) &= \frac{1}{4a_1 a_2}\int_{-a_1}^{a_1} dx_1 \int_{-a_2}^{a_2} dx_2 \, f(\mathbf{x}) \\[2mm]
&= \frac{1}{4a_1 a_2}\mathrm{Erf}\left[\frac{a_2}{\sigma_2\sqrt{2(1-\rho^2)}}\right]\mathrm{Erf}\left[\frac{a_1}{\sigma_1\sqrt{2}}\right] \\[2mm]
&\quad - \frac{\rho^2 e^{-\frac{a_2^2}{2\sigma_2^2(1-\rho^2)}}}{2\sigma_1\sigma_2(1-\rho^2)\sqrt{2\pi(1-\rho^2)}}\frac{\mathrm{Erf}\left[\frac{a_1}{\sigma_1\sqrt{2}}\right]}{2a_1} + \frac{\rho^2 e^{-\frac{a_2^2}{2\sigma_2^2(1-\rho^2)} - \frac{a_1^2}{2\sigma_1^2}}}{4\pi\sigma_1^2\sigma_2\sqrt{1-\rho^2}}.
\end{aligned} \tag{30}$$

Note that in the limit of no cross-correlations, $\rho \to 0$, the integral factorizes and we can write an exact expression for the evidence,

$$
E(a_1, a_2) = \frac{1}{4a_1 a_2} \frac{1}{2\pi\sigma_1\sigma_2} \int_{-a_1}^{a_1} dx_1 \int_{-a_2}^{a_2} dx_2 \, e^{-\frac{x_1^2}{2\sigma_1^2} - \frac{x_2^2}{2\sigma_2^2}} \tag{31}
$$

$$
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} du_1 \int_{-\infty}^{\infty} du_2 \, \frac{\sin a_1 u_1}{a_1 u_1} \frac{\sin a_2 u_2}{a_2 u_2} e^{-\frac{1}{2}\sigma_1^2 u_1^2 - \frac{1}{2}\sigma_2^2 u_2^2} \tag{32}
$$

$$
= \frac{1}{4a_1 a_2} \mathrm{Erf}\left[\frac{a_1}{\sigma_1\sqrt{2}}\right] \mathrm{Erf}\left[\frac{a_2}{\sigma_2\sqrt{2}}\right]. \tag{33}
$$

It happens, however, that even in the presence of cross-correlations, if the prior is wide ($a_i \geq 2\sigma_i$), then the terms proportional to exponentials are negligible and the evidence becomes, to a very good approximation,

$$
E(a_1, a_2) = \frac{1}{4a_1 a_2} \mathrm{Erf}\left[\frac{a_2}{\sigma_2\sqrt{2(1-\rho^2)}}\right] \mathrm{Erf}\left[\frac{a_1}{\sigma_1\sqrt{2}}\right]. \tag{34}
$$

Moreover, in that case, the error functions are approximately given by 1.

n **variables**. Suppose we have n correlated variables, $\mathbf{x} = (x_1, \ldots, x_n)$, with covariance matrix

$$
C_n = \begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1n} \\ C_{12} & C_{22} & \ldots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \ldots & C_{nn} \end{pmatrix}. \tag{35}
$$

In this case, the probability distribution function can be expressed as

$$
f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C_n}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1} \mathbf{x}\right], \tag{36}
$$

which has the property that marginalizing over the last variable, x_n, we obtain a correlated probability distribution function for the $n-1$ variables, $\mathbf{x} = (x_1, \ldots, x_{n-1})$,

$$
f(\mathbf{x}) = \frac{1}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1} \mathbf{x}\right], \tag{37}
$$

where the C_{n-1} covariance matrix is given by Equation (35) without the last column and last row.

We will now evaluate the evidence Equation (22) for this multivariate Gaussian, starting with the integration over the last variable, x_n,

$$
\frac{1}{2a_n} \int_{-a_n}^{a_n} dx_n \, f(\mathbf{x}) = \frac{1}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1} \mathbf{x}\right]
$$

$$
\times \left\{ \frac{1}{2a_n} \mathrm{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \mathcal{O}\left(e^{-\frac{a_n^2 \det C_{n-1}}{2 \det C_n}}\right) \right\}. \tag{38}
$$

Integrating now over the next variable, x_{n-1}, we find

$$\frac{1}{4a_n a_{n-1}} \int_{-a_n}^{a_n} dx_n \int_{-a_{n-1}}^{a_{n-1}} dx_{n-1}\, f(\mathbf{x}) = \frac{1}{(2\pi)^{(n-2)/2}\sqrt{\det C_{n-2}}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-2}^{-1}\mathbf{x}\right]$$

$$\times \left\{ \frac{1}{4a_n a_{n-1}} \mathrm{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] \mathrm{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-2}}{\det C_{n-1}}}\right] + \mathcal{O}\left(e^{-\frac{a_n^2 \det C_{n-1}}{2\det C_n}}\right) \right\}. \tag{39}$$

Continuing the integration over the priors, we end up with the evidence for the *n*-dimensional distribution,

$$\begin{aligned} E(a_1,\ldots,a_n) &= \frac{1}{\prod_{p=1}^{n} 2a_p} \int_{-a_1}^{a_1}\cdots\int_{-a_n}^{a_n} d^n\mathbf{x}\, f(\mathbf{x}) \\[2mm] &= \prod_{p=1}^{n} \frac{1}{2a_p}\mathrm{Erf}\left[\frac{a_p}{\sqrt{2}}\sqrt{\frac{\det C_{p-1}}{\det C_p}}\right] + \mathcal{O}\left(\exp\left[-\sum_{p=1}^{n} \frac{a_p^2 \det C_{p-1}}{2\det C_p}\right]\right), \end{aligned} \tag{40}$$

where the covariance matrices C_p are constructed as above, by eliminating the $n-p$ last rows and columns, until we end up with $C_0 \equiv 1$. Note that the approximation is very good whenever $\sum_{p=1}^{n}(a_p^2 \det C_{p-1})/(2\det C_p) \gg 1$, which is often the case. Note also that we recover the previous result Equation (34) for the particular case $n=2$.

In the limit that the cross-correlation between the *n* variables vanishes, the evidence (40) reduces to the exact result

$$E(a_1,\ldots,a_n) = \prod_{p=1}^{n} \frac{1}{2a_p}\mathrm{Erf}\left[\frac{a_p}{\sigma_p\sqrt{2}}\right]. \tag{41}$$

Note that the evidence Equation (40) correctly reflects the limit in which we eliminate the need for a new variable x_n, by making its prior vanish,

$$\lim_{a_n \to 0} E(a_1,\ldots,a_n) = E(a_1,\ldots,a_{n-1})\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}, \tag{42}$$

and thus we recover in that limit a properly normalized distribution, $f(x_1,\ldots,x_n) \to f(x_1,\ldots,x_{n-1})$, while the inspection of the likelihood function alone would not have been able to give a reasonable answer.

On the other hand, in the case that our theoretical prejudice cannot assign a concrete prior to a given variable, we see that the evidence decreases as $1/2a$ as *a* increases. Therefore, the Bayesian evidence seems to be a very good discriminator between theoretical priors, and penalizes including too many parameters, a la Occam's razor.

3.2. Uncentred Priors

It is unlikely that the priors will actually be centred on the mean of the distribution, as the priors are not supposed to know what the data will tell us. We therefore need to generalize the above for uncentred priors. We continue to assume that the priors are top hats.

We also continue to assume for the moment that the probability distribution is well-approximated by a Gaussian with mean value μ. We will then use displaced variables $\tilde{x}_i = x_i - \mu_i$, and write the Gaussian distribution function as in Equation (36). The normalized top-hat prior is now uncentered with respect to the mean value,

$$\pi(\tilde{x};a,b) \equiv \begin{cases} (a+b)^{-1} & -a < \tilde{x} < b, \\ 0 & \text{otherwise}. \end{cases} \tag{43}$$

For a single variable, the result is exact,

$$E(a; b) = \int_{-\infty}^{\infty} dx \, f(x) \, \pi(x; a, b) = \frac{1}{2a + 2b} \left(\text{Erf}\left[\frac{a}{\sigma\sqrt{2}}\right] + \text{Erf}\left[\frac{b}{\sigma\sqrt{2}}\right] \right). \tag{44}$$

where we are integrating over the displaced variable $\tilde{x}$, from now on renamed as x. Note that we recover the result Equation (24) for the centred prior case in the limit $b \to a$.

For two variables, with distribution function Equation (26), the uncentred Bayesian evidence is

$$E(a_1, a_2; b_1, b_2) = \frac{1}{(a_1 + b_1)(a_2 + b_2)} \int_{-a_1}^{b_1} dx_1 \int_{-a_2}^{b_2} dx_2 \, f(x_1, x_2) \tag{45}$$

$$= \frac{1}{(2a_1 + 2b_1)(2a_2 + 2b_2)} \left\{ \left(\text{Erf}\left[\frac{a_1}{\sigma_1\sqrt{2}}\right] + \text{Erf}\left[\frac{b_1}{\sigma_1\sqrt{2}}\right] \right) \right. \tag{46}$$

$$\times \left(\text{Erf}\left[\frac{a_2}{\sigma_2\sqrt{2(1-\rho^2)}}\right] + \text{Erf}\left[\frac{b_2}{\sigma_2\sqrt{2(1-\rho^2)}}\right] \right)$$

$$\left. - \frac{\rho}{2\pi\sqrt{1-\rho^2}} \left(e^{-\frac{a_1^2}{2\sigma_1^2}} - e^{-\frac{b_1^2}{2\sigma_1^2}} \right) \left(e^{-\frac{a_2^2}{2\sigma_2^2(1-\rho^2)}} + e^{-\frac{b_2^2}{2\sigma_2^2(1-\rho^2)}} \right) \right\}$$

The evidence for the multiple-variable case Equation (36) is

$$E(\mathbf{a}, \mathbf{b}) = \int_{-\infty}^{\infty} d^n\mathbf{x} \, f(\mathbf{x}) \prod_{i=1}^{n} \pi(x_i; a_i, b_i) = \prod_{i=1}^{n} (a_i + b_i)^{-1} \int_{-a_1}^{b_1} d\tilde{x}_1 \cdots \int_{-a_n}^{b_n} d\tilde{x}_n \, f(\tilde{\mathbf{x}}). \tag{47}$$

Let us now evaluate it for the multivariate Gaussian Equation (36), starting with the integration over the last variable, x_n,

$$\frac{1}{a_n + b_n} \int_{-a_n}^{b_n} dx_n \, f(\mathbf{x}) = \frac{1}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right] \frac{1}{(2a_n + 2b_n)}$$

$$\times \left\{ \text{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \text{Erf}\left[\frac{b_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \mathcal{O}\left(e^{-\frac{a_n^2 \det C_{n-1}}{2\det C_n}} + e^{-\frac{b_n^2 \det C_{n-1}}{2\det C_n}}\right) \right\} \tag{48}$$

Integrating now over the next variable, x_{n-1}, we find

$$\frac{1}{(a_n + b_n)(a_{n-1} + b_{n-1})} \int_{-a_n}^{b_n} dx_n \int_{-a_{n-1}}^{b_{n-1}} dx_{n-1} \, f(\mathbf{x}) =$$

$$\frac{1}{(2\pi)^{(n-2)/2}\sqrt{\det C_{n-2}}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-2}^{-1}\mathbf{x}\right] \frac{1}{(2a_n + 2b_n)(2a_{n-1} + 2b_{n-1})} \tag{49}$$

$$\times \left\{ \left(\text{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \text{Erf}\left[\frac{b_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] \right) \tag{50}$$

$$\times \left(\text{Erf}\left[\frac{a_{n-1}}{\sqrt{2}}\sqrt{\frac{\det C_{n-2}}{\det C_{n-1}}}\right] + \text{Erf}\left[\frac{b_{n-1}}{\sqrt{2}}\sqrt{\frac{\det C_{n-2}}{\det C_{n-1}}}\right] \right) \tag{51}$$

$$\left. + \mathcal{O}\left(e^{-\frac{a_n^2 \det C_{n-1}}{2\det C_n}} + e^{-\frac{b_n^2 \det C_{n-1}}{2\det C_n}}\right) \times \left(e^{-\frac{a_{n-1}^2 \det C_{n-2}}{2\det C_{n-1}}} + e^{-\frac{b_{n-1}^2 \det C_{n-2}}{2\det C_{n-1}}}\right) \right\}.$$

Continuing the integration over the priors, we end up with the evidence for the n-dimensional distribution,

$$
\begin{aligned}
E(\mathbf{a}, \mathbf{b}) \;=\;& \frac{1}{\prod_{p=1}^{n}(a_p + b_p)} \int_{-a_1}^{b_1} \cdots \int_{-a_n}^{b_n} d^n\mathbf{x}\, f(\mathbf{x}) \\[2mm]
=\;& \prod_{p=1}^{n} \frac{1}{(2a_p + 2b_p)} \left(\mathrm{Erf}\left[\frac{a_p}{\sqrt{2}} \sqrt{\frac{\det C_{p-1}}{\det C_p}} \right] + \mathrm{Erf}\left[\frac{b_p}{\sqrt{2}} \sqrt{\frac{\det C_{p-1}}{\det C_p}} \right] \right) \\[2mm]
& + \mathcal{O}\left(\prod_{p=1}^{n} \left[\exp\left(- \frac{a_p^2 \det C_{p-1}}{2 \det C_p} \right) + \exp\left(- \frac{b_p^2 \det C_{p-1}}{2 \det C_p} \right) \right] \right),
\end{aligned}
\tag{52}
$$

where the covariance matrices C_p are constructed as above, by eliminating the $n - p$ last rows and columns, until $C_0 \equiv 1$. Note that the approximation is very good whenever the exponents are large, $\sum_{p=1}^{n}(a_p^2 \det C_{p-1})/(2 \det C_p) \gg 1$, which is often the case. Note also that we recover the expression of the evidence for the centred priors Equation (40) in the limit $b \to a$.

Let us now evaluate the evidence for a distribution normalized to the maximum of the likelihood distribution,

$$
f(\mathbf{x}) = \mathcal{L}_{\max} \exp\left[-\frac{1}{2} \mathbf{x}^T C_n^{-1} \mathbf{x} \right]
\tag{53}
$$

In this case, the evidence is given by Equation (52), multiplied by a factor $\mathcal{L}_{\max} \times (2\pi)^{n/2} \sqrt{\det C_n}$ from the normalization. We can then evaluate the logarithm of the evidence, ignoring the exponentially small corrections, as

$$
\begin{aligned}
\ln E \;=\;& \ln \mathcal{L}_{\max} + \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln \det C_n - \sum_{p=1}^{n} \ln(2a_p + 2b_p) \\[2mm]
& + \sum_{p=1}^{n} \ln\left(\mathrm{Erf}\left[\frac{a_p}{\sqrt{2}} \sqrt{\frac{\det C_{p-1}}{\det C_p}} \right] + \mathrm{Erf}\left[\frac{b_p}{\sqrt{2}} \sqrt{\frac{\det C_{p-1}}{\det C_p}} \right] \right).
\end{aligned}
\tag{54}
$$

Uncorrelated case. Suppose we have a multivariate Gaussian distribution without correlations between variables, i.e., $C_{ij} = \sigma_i^2 \delta_{ij}$ is a diagonal matrix; then the evidence reads exactly,

$$
E(\mathbf{a}, \mathbf{b}) = \frac{1}{\prod_{p=1}^{n}(a_p + b_p)} \int_{-a_1}^{b_1} \cdots \int_{-a_n}^{b_n} d^n\mathbf{x}\, f(\mathbf{x}) = \prod_{p=1}^{n} \frac{1}{2(a_p + b_p)} \left(\mathrm{Erf}\left[\frac{a_p}{\sigma_p \sqrt{2}} \right] + \mathrm{Erf}\left[\frac{b_p}{\sigma_p \sqrt{2}} \right] \right),
\tag{55}
$$

where σ_p are the dispersions of each variable $\tilde{x}_p$, and thus the logarithm of the evidence becomes

$$
\ln E = \ln \mathcal{L}_{\max} + \frac{n}{2} \ln(2\pi) + \sum_{p=1}^{n} \ln \sigma_p - \sum_{p=1}^{n} \ln(2a_p + 2b_p) + \sum_{p=1}^{n} \ln\left(\mathrm{Erf}\left[\frac{a_p}{\sigma_p \sqrt{2}} \right] + \mathrm{Erf}\left[\frac{b_p}{\sigma_p \sqrt{2}} \right] \right)
\tag{56}
$$

Laplace approximation. The Laplacian approximation to the evidence assumes the distribution is a correlated Gaussian, and that the priors are large enough so that the whole distribution fits easily inside them, in which case the error functions are approximately in unity and do not contribute to the evidence; from Equation (54) we now have

$$
\ln E = \ln \mathcal{L}_{\max} + \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln \det C_n - \sum_{p=1}^{n} \ln \Delta\theta_p,
\tag{57}
$$

where $\Delta\theta_p = a_p + b_p$ is the parameter interval associated to the prior. In the next section we will compare the different approximations.

4. Non-Gaussian Corrections

The advantage of this method is that one can perform a systematic computation of the evidence of a given model with its own priors, given an arbitrary set of moments of the distribution. Here we will consider the first two beyond the covariance matrix, i.e., the skewness and kurtosis terms, see Equation (9).

4.1. Skewness

Let us start with the first correction to the Gaussian approximation, the trilinear term B_{ijk}. For this, we write the generating functional (9) as

$$\phi(\mathbf{u}) = \exp\left[i\,\mu_i u_i - \frac{1}{2!}\,C_{ij}\,u_i u_j - \frac{i}{3!}\,B_{ijk}\,u_i u_j u_k\right]. \tag{58}$$

By performing a change of variable, $u_i = y_i - i\,C_{ik}^{-1}(x_k - \mu_k)$, we can evaluate the Fourier transform integral and obtain the properly-normalized probability distribution function

$$
\begin{aligned}
f(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det C_n}}\exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right] \\
&\quad \times \left(1 - \frac{1}{2}B_{ijk}\,C_{ij}^{-1}C_{kl}^{-1}\,x_l + \frac{1}{6}B_{ijk}\,C_{il}^{-1}C_{jm}^{-1}C_{kn}^{-1}\,x_l x_m x_n\right),
\end{aligned} \tag{59}
$$

where x_k are the displaced coordinates $(x_k - \mu_k)$. This skewed distribution function satisfies

$$\langle x_i \rangle = 0, \qquad \langle x_i x_j \rangle = C_{ij}, \qquad \langle x_i x_j x_k \rangle = B_{ijk}, \qquad \langle x_i x_j x_k x_l \rangle = 0, \quad \dots \tag{60}$$

as can be confirmed by direct evaluation. Let us now compute the evidence Equation (22) for this skewed model. Since the extra terms in the parenthesis of Equation (59) are both odd functions of x, when integrating over an even range like that of the centred top-hat prior Equation (21), their contribution to the evidence vanish, and thus the final evidence for the skewed model does not differ from that of the Gaussian model Equation (40). In case the prior is off-centred with respect to the mean, e.g., in Equation (43), then the contribution of the odd terms to the evidence would not vanish. Let us evaluate their contribution.

For a single variable ($n = 1$), the correctly normalized likelihood function can be written as

$$f(x) = \frac{e^{-x^2/2\sigma^2}}{\sigma\sqrt{2\pi}}\left(1 - \frac{B\,x}{2\sigma^4} + \frac{B\,x^3}{6\sigma^6}\right),$$

satisfying $\langle x \rangle = 0$, $\langle x^2 \rangle = \sigma^2$, $\langle x^3 \rangle = B$, and the Bayesian integral can be computed exactly as

$$E(a,b) = \frac{1}{2a+2b}\left(\mathrm{Erf}\left[\frac{a}{\sigma\sqrt{2}}\right] + \mathrm{Erf}\left[\frac{b}{\sigma\sqrt{2}}\right]\right) - \frac{B\sigma^{-3}}{6\sqrt{2\pi}}\left[\left(1 - \frac{a^2}{\sigma^2}\right)e^{-\frac{a^2}{2\sigma^2}} - \left(1 - \frac{b^2}{\sigma^2}\right)e^{-\frac{b^2}{2\sigma^2}}\right]\frac{1}{a+b}. \tag{61}$$

Note that for even (centred) priors, with $b = a$, the evidence reduces to Equation (24).

For an arbitrary number of variables the computation is more complicated. Let us start with the n-th variable and, in order to compute the integral, let us define the auxiliary function

$$g(\lambda) = \int_{-a_n}^{b_n} dx_n \, x_n \, \frac{\exp\left[-\frac{\lambda}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]}{(2\pi)^{n/2}\sqrt{\det C_n}} = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \times$$

$$\times \frac{1}{\lambda\sqrt{2\pi}}\left(\exp\left[-\frac{\lambda a_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] - \exp\left[-\frac{\lambda b_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right]\right), \tag{62}$$

such that, using $\mathrm{Erf}'[x] = \frac{2}{\sqrt{\pi}}e^{-x^2}$,

$$-2g'(\lambda = 1) = \int_{-a_n}^{b_n} dx_n \, x_n \, \frac{(\mathbf{x}^T C_n^{-1}\mathbf{x})\exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]}{(2\pi)^{n/2}\sqrt{\det C_n}} = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \times$$

$$\times \frac{1}{\sqrt{2\pi}}\left\{\left(2 + a_n^2\frac{\det C_{n-1}}{\det C_n}\right)\exp\left[-\frac{a_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] - \left(2 + b_n^2\frac{\det C_{n-1}}{\det C_n}\right)\exp\left[-\frac{b_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right]\right\}. \tag{63}$$

Therefore, with the use of Equation (63), the integral of the skewness-corrected distribution function Equation (59) over the x_n uncentred prior becomes

$$\int_{-a_n}^{b_n} dx_n \, f(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}}\left\{\frac{1}{2}\left(\mathrm{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \mathrm{Erf}\left[\frac{b_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right]\right)\right.$$

$$\left. - \frac{1}{6}B_{ijn}C_{ij}^{-1}\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\left[\left(1 - a_n^2\frac{\det C_{n-1}}{\det C_n}\right)e^{-\frac{a_n^2\det C_{n-1}}{2\det C_n}} - \left(1 - b_n^2\frac{\det C_{n-1}}{\det C_n}\right)e^{-\frac{b_n^2\det C_{n-1}}{2\det C_n}}\right]\right\}. \tag{64}$$

Let us define two new functions,

$$E_i(a_i, b_i) = \frac{1}{2}\left(\mathrm{Erf}\left[\frac{a_i}{\sqrt{2}}\sqrt{\frac{\det C_{i-1}}{\det C_i}}\right] + \mathrm{Erf}\left[\frac{b_i}{\sqrt{2}}\sqrt{\frac{\det C_{i-1}}{\det C_i}}\right]\right), \tag{65}$$

$$F_i(a_i, b_i) = \frac{1}{6\sqrt{2\pi}}\sqrt{\frac{\det C_{i-1}}{\det C_i}}\left[\left(1 - a_i^2\frac{\det C_{i-1}}{\det C_i}\right)e^{-\frac{a_i^2\det C_{i-1}}{2\det C_i}} - \left(1 - b_i^2\frac{\det C_{i-1}}{\det C_i}\right)e^{-\frac{b_i^2\det C_{i-1}}{2\det C_i}}\right].$$

Integrating iteratively over $x_{n-1}, \ldots, x_1$, we end up with the Bayesian evidence for the third-order-corrected probability distribution function $f(\mathbf{x})$,

$$E(\mathbf{a}, \mathbf{b}) = \prod_{p=1}^{n}\frac{E_p(a_p, b_p)}{(a_p + b_p)}\left[1 - \sum_{k=1}^{n}B_{ijk}C_{ij}^{-1}\frac{F_k(a_k, b_k)}{E_k(a_k, b_k)}\right]. \tag{66}$$

Unless $B_{ijk}C_{ij}^{-1}$ is very large, the correction to the error function is exponentially suppressed, and we do not expect significant departures from the Gaussian case Equation (40). Note also that if the prior is symmetrical, it is easy to see that the skewness part of the integral vanishes, $F_k(a_k, b_k) \to 0$, as can be checked explicitly by taking $b_k \to a_k$.

4.2. Kurtosis

The next correction beyond skewness is the fourth-order moment or kurtosis, given by the D_{ijkl} term in Equation (9). Let us ignore for the moment the third-order skewness and write

$$\phi(\mathbf{u}) = \exp\left[i\,\mu_i u_i - \frac{1}{2!}C_{ij}\,u_i u_j + \frac{1}{4!}D_{ijkl}\,u_i u_j u_k u_l\right]. \tag{67}$$

By performing the same change of variables, $u_i = y_i - iC_{ik}^{-1}(x_k - \mu_k)$, we can now compute the Fourier transform and obtain the properly normalized probability distribution function

$$
f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C_n}} \exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]\left(1 + \frac{1}{8}D_{ijkl}\,C_{ij}^{-1}C_{kl}^{-1}\right.
$$

$$
\left. - \frac{1}{4}D_{ijkl}\,C_{ij}^{-1}C_{km}^{-1}C_{ln}^{-1}\,x_m x_n + \frac{1}{24}D_{ijkl}\,C_{im}^{-1}C_{jn}^{-1}C_{kp}^{-1}C_{lq}^{-1}\,x_m x_n x_p x_q\right). \quad (68)
$$

Performing the integrals, it is easy to see that this distribution satisfies

$$
\langle x_i x_j \rangle = C_{ij}, \qquad \langle x_i x_j x_k x_l \rangle = D_{ijkl} + C_{ij}C_{kl} + C_{ik}C_{jl} + C_{il}C_{jk}, \qquad \dots \quad (69)
$$

Note that in order for the new likelihood distribution (68) to be positive definite, it is required that $D_{ijkl}C_{ij}^{-1}C_{kl}^{-1} < 4$, and if we impose that there is only one maximum at the centre, then it must satisfy $D_{ijkl}C_{ij}^{-1}C_{kl}^{-1} < 2$. These conditions impose bounds on the maximum possible deviation of the evidence from a that of a Gaussian.

Let us now compute the evidence Equation (22) for this kurtosis model. The extra terms in the parenthesis of Equation (68) are both even functions of x, and we cannot ignore them, even for centred priors.

For a single variable ($n = 1$), the correctly normalized likelihood function can be written as

$$
f(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\left(1 + \frac{D}{8\sigma^4} - \frac{D\,x^2}{4\sigma^6} + \frac{D\,x^4}{24\sigma^8}\right),
$$

satisfying $\langle x \rangle = 0$, $\langle x^2 \rangle = \sigma^2$, $\langle x^3 \rangle = 0$, $\langle x^4 \rangle = D + 3\sigma^4$, etc. The Bayesian integral can be computed exactly as

$$
E(a,b) = \frac{1}{2a+2b}\left(\text{Erf}\left[\frac{a}{\sigma\sqrt{2}}\right] + \text{Erf}\left[\frac{b}{\sigma\sqrt{2}}\right]\right) + \frac{D\sigma^{-4}}{8\sqrt{2\pi}}\left(\frac{a}{\sigma}\left(1 - \frac{a^2}{3\sigma^2}\right)e^{-\frac{a^2}{2\sigma^2}} + \frac{b}{\sigma}\left(1 - \frac{b^2}{3\sigma^2}\right)e^{-\frac{b^2}{2\sigma^2}}\right)\frac{1}{a+b}. \quad (70)
$$

For an arbitrary number of variables, the computation is again much more complicated. Let us start with the n-th variable and, in order to compute the first integral, let us define a new auxiliary function

$$
h(\lambda) = \int_{-a_n}^{b_n} dx_n \frac{\exp\left[-\frac{\lambda}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]}{(2\pi)^{n/2}\sqrt{\det C_n}} = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \times
$$

$$
\times \frac{1}{2\sqrt{\lambda}}\left(\text{Erf}\left[\frac{a_n\sqrt{\lambda}}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \text{Erf}\left[\frac{b_n\sqrt{\lambda}}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right]\right), \quad (71)
$$

such that,

$$
-2h'(\lambda=1) = \int_{-a_n}^{b_n} dx_n \frac{(\mathbf{x}^T C_n^{-1}\mathbf{x})\exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]}{(2\pi)^{n/2}\sqrt{\det C_n}} = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \times
$$

$$
\times \left\{\frac{1}{2}\left(\text{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \text{Erf}\left[\frac{b_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right]\right)\right. \quad (72)
$$

$$
\left. - \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\left(a_n \exp\left[-\frac{a_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] + b_n \exp\left[-\frac{b_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right]\right)\right\}.
$$

$$
4h''(\lambda = 1) = \int_{-a_n}^{b_n} dx_n \frac{(\mathbf{x}^T C_n^{-1}\mathbf{x})^2 \exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]}{(2\pi)^n \sqrt{\det C_n}} = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}} \times
$$

$$
\times \left\{ \frac{3}{2}\left(\mathrm{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \mathrm{Erf}\left[\frac{b_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] \right) \right. \tag{73}
$$

$$
-\frac{3}{\sqrt{2\pi}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\left(a_n \exp\left[-\frac{a_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] + b_n \exp\left[-\frac{b_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] \right)
$$

$$
\left. -\frac{a_n^2}{\sqrt{2\pi}}\left(\frac{\det C_{n-1}}{\det C_n}\right)^{3/2}\left(a_n \exp\left[-\frac{a_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] + b_n \exp\left[-\frac{b_n^2}{2}\frac{\det C_{n-1}}{\det C_n}\right] \right) \right\}.
$$

Therefore, with the use of Equations (72) and (73), the integral of the kurtosis-corrected distribution function (68) over the x_n prior becomes

$$
\int_{-a_n}^{b_n} dx_n\, f(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}\mathbf{x}^T C_{n-1}^{-1}\mathbf{x}\right]}{(2\pi)^{(n-1)/2}\sqrt{\det C_{n-1}}}\left\{ \frac{1}{2}\left(\mathrm{Erf}\left[\frac{a_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] + \mathrm{Erf}\left[\frac{b_n}{\sqrt{2}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\right] \right) + \right. \tag{74}
$$

$$
\left. +\frac{1}{8}D_{ijkl}\,C_{ij}^{-1}C_{kl}^{-1}\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\det C_{n-1}}{\det C_n}}\left[a_n\left(1-\frac{a_n^2}{3}\frac{\det C_{n-1}}{\det C_n}\right)e^{-\frac{a_n^2}{2}\frac{\det C_{n-1}}{\det C_n}} + b_n\left(1-\frac{b_n^2}{3}\frac{\det C_{n-1}}{\det C_n}\right)e^{-\frac{b_n^2}{2}\frac{\det C_{n-1}}{\det C_n}} \right] \right\}.
$$

We can now define a new function

$$
G_i(a_i, b_i) = \frac{1}{8\sqrt{2\pi}}\sqrt{\frac{\det C_{i-1}}{\det C_i}}\left[a_i\left(1-\frac{a_i^2}{3}\frac{\det C_{i-1}}{\det C_i}\right)e^{-\frac{a_i^2}{2}\frac{\det C_{i-1}}{\det C_i}} - b_i\left(1-\frac{b_i^2}{3}\frac{\det C_{i-1}}{\det C_i}\right)e^{-\frac{b_i^2}{2}\frac{\det C_{i-1}}{\det C_i}} \right]. \tag{75}
$$

Integrating iteratively over $x_{n-1}, \ldots, x_1$, we end up with the Bayesian evidence for the fourth-order-corrected probability distribution function $f(\mathbf{x})$,

$$
E(\mathbf{a}, \mathbf{b}) = \prod_{p=1}^{n}\frac{E_p(a_p, b_p)}{(a_p + b_p)}\left[1 + D_{ijkl}\,C_{ij}^{-1}C_{kl}^{-1}\sum_{m=1}^{n}\frac{G_m(a_m, b_m)}{E_m(a_m, b_m)} \right]. \tag{76}
$$

so, unless $D_{ijkl}\,C_{ij}^{-1}C_{kl}^{-1}$ is very large, the correction to the error function is exponentially suppressed, and we do not expect significant departures from the Gaussian case, Equation (40).

In order to compare models it is customary to compute the logarithm of the evidence. Let us assume that we are given a likelihood distribution function normalized by the maximum likelihood, and with corrections up to the fourth order,

$$
f(\mathbf{x}) = \mathcal{L}_{\max}\exp\left[-\frac{1}{2}\mathbf{x}^T C_n^{-1}\mathbf{x}\right]\left(1+\frac{1}{8}D_{ijkl}\,C_{ij}^{-1}C_{kl}^{-1}\right)^{-1}\left(1-\frac{1}{2}B_{ijk}\,C_{ij}^{-1}C_{kl}^{-1}x_l + \frac{1}{6}B_{ijk}\,C_{il}^{-1}C_{jm}^{-1}C_{kn}^{-1}x_l x_m x_n\right.
$$

$$
\left. +\frac{1}{8}D_{ijkl}\,C_{ij}^{-1}C_{kl}^{-1} - \frac{1}{4}D_{ijkl}\,C_{ij}^{-1}C_{km}^{-1}C_{ln}^{-1}x_m x_n + \frac{1}{24}D_{ijkl}\,C_{im}^{-1}C_{jn}^{-1}C_{kp}^{-1}C_{lq}^{-1}x_m x_n x_p x_q\right). \tag{77}
$$

Note that it is normalized so that the maximum corresponds to the mean-centred distribution, i.e., $\mathbf{x} = 0$. In this case, the evidence of the normalized distribution is given by

$$E(\mathbf{a}, \mathbf{b}) = \mathcal{L}_{\max} (2\pi)^{n/2} \sqrt{\det C_n} \left(1 + \frac{1}{8} D_{ijkl} C_{ij}^{-1} C_{kl}^{-1} \right)^{-1} \times \tag{78}$$

$$\prod_{p=1}^{n} \frac{E_p(a_p, b_p)}{(a_p + b_p)} \left[1 - \sum_{k=1}^{n} B_{ijk} C_{ij}^{-1} \frac{F_k(a_k, b_k)}{E_k(a_k, b_k)} + D_{ijkl} C_{ij}^{-1} C_{kl}^{-1} \sum_{m=1}^{n} \frac{G_m(a_m, b_m)}{E_m(a_m, b_m)} \right].$$

We can then evaluate the logarithm of the evidence by

$$\ln E = \ln \mathcal{L}_{\max} + \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln \det C_n - \ln \left(1 + \frac{1}{8} D_{ijkl} C_{ij}^{-1} C_{kl}^{-1} \right) - \sum_{p=1}^{n} \ln(2a_p + 2b_p)$$

$$+ \sum_{p=1}^{n} \ln \left(\mathrm{Erf} \left[\frac{a_p}{\sqrt{2}} \sqrt{\frac{\det C_{p-1}}{\det C_p}} \right] + \mathrm{Erf} \left[\frac{b_p}{\sqrt{2}} \sqrt{\frac{\det C_{p-1}}{\det C_p}} \right] \right) \tag{79}$$

$$+ \ln \left(1 - \sum_{k=1}^{n} B_{ijk} C_{ij}^{-1} \frac{F_k(a_k, b_k)}{E_k(a_k, b_k)} + D_{ijkl} C_{ij}^{-1} C_{kl}^{-1} \sum_{m=1}^{n} \frac{G_m(a_m, b_m)}{E_m(a_m, b_m)} \right).$$

Note that the condition $D_{ijkl} C_{ij}^{-1} C_{kl}^{-1} < 2$ constrains the maximum amount that the kurtosis corrections can contribute to the evidence.

Uncorrelated case. In the case where the likelihood distribution has no correlations among the different variables, the exact expression for the Bayesian evidence is

$$\ln E = \ln \mathcal{L}_{\max} + \frac{n}{2} \ln(2\pi) + \sum_{p=1}^{n} \ln \sigma_p - \sum_{p=1}^{n} \ln(2a_p + 2b_p) + \sum_{p=1}^{n} \ln \left(\mathrm{Erf} \left[\frac{a_p}{\sigma_p \sqrt{2}} \right] + \mathrm{Erf} \left[\frac{b_p}{\sigma_p \sqrt{2}} \right] \right) \tag{80}$$

$$- \ln \left(1 + \frac{1}{8} D_{iijj} \sigma_i^{-2} \sigma_j^{-2} \right) + \ln \left(1 - \sum_{k=1}^{n} B_{iik} \sigma_k^{-2} \frac{F_k(a_k, b_k)}{E_k(a_k, b_k)} + D_{iijj} \sigma_i^{-2} \sigma_j^{-2} \sum_{m=1}^{n} \frac{G_m(a_m, b_m)}{E_m(a_m, b_m)} \right),$$

where σ_p are the corresponding dispersions of variables x_p, and the functions E_i, F_i and G_i are the corresponding limiting functions of Equations (65) and (75) for uncorrelated matrices.

5. Model Comparison

Finally we turn to specific applications of the formalism discussed above. Initially we will carry out some toy model tests of its performance, and then examine real cosmological applications for which we previously obtained results by thermodynamic integration [12].

5.1. A Baby-Toy Model Comparison

We begin with a very simple two-dimensional toy model. The purpose of this section is to illustrate the ineffectiveness of the thermodynamic integration and to give an indication of the performance of the method we propose here. In addition, the two-dimensional model is simple enough to allow a brute-force direct numerical integration of evidence allowing us to check the accuracy at the same time. We use the following two forms of likelihood:

$$\mathcal{L}_g(x, y) = \exp \left[-\frac{2x^2 - 2(y-1)^2 - xy}{2} \right] \tag{81}$$

$$\mathcal{L}_{ng}(x, y) = \exp \left[-\frac{2x^2 - 2(y-1)^2 - xy}{2} \right] + \exp \left[-\frac{2x^2 - 2y^2 - 3xy}{2} \right] \tag{82}$$

The subscripts g and ng indicate the Gaussian and non-Gaussian cases, respectively.

Firstly, we calculate the evidence by the analytical method using Equations (56) and (80) and covariance matrices inferred from sampling the likelihood using the vanilla Metropolis–Hastings algorithm with fixed proposal widths. Chains ranging from a few to several millions of samples were used. We also calculate evidence using thermodynamic algorithm explained in ref. [12]. Again, we vary algorithm parameters to obtain evidence values of varying accuracy. The resulting evidence as a function of the number of likelihood evaluations is plotted in the Figure 1, together with the correct value inferred by direct numerical integration. The number of likelihood evaluations is crucial as this is the time-limiting step in the cosmological parameter estimation and model comparison exercises. The results are what could have been anticipated. We note that the size of the prior does not seem to be of crucial importance. This is comforting, given that the analytical method requires the knowledge of the true covariance information, while we can only supply a covariance matrix estimated from the prior-truncated likelihood. We also note that the thermodynamic integration converges to the correct value in all cases. However, it does so after very many likelihood evaluations; typically about a million or so even for a two-dimensional problem. The analytical method already becomes limited by systematics by the ten-thousand samples. For the Gaussian case, there is no systematic by construction, while the non-Gaussian case suffers a systematic of about 0.1 in ln E. The non-Gaussian correction reduces the error by about half and thus correctly estimates the uncertainty associated with the purely Gaussian approximation. In the case of wide priors, the only non-Gaussian correction of appreciable size is the $\ln(1 + D_{ijkl}C_{ij}^{-1}C_{kl}^{-1}/8)$.
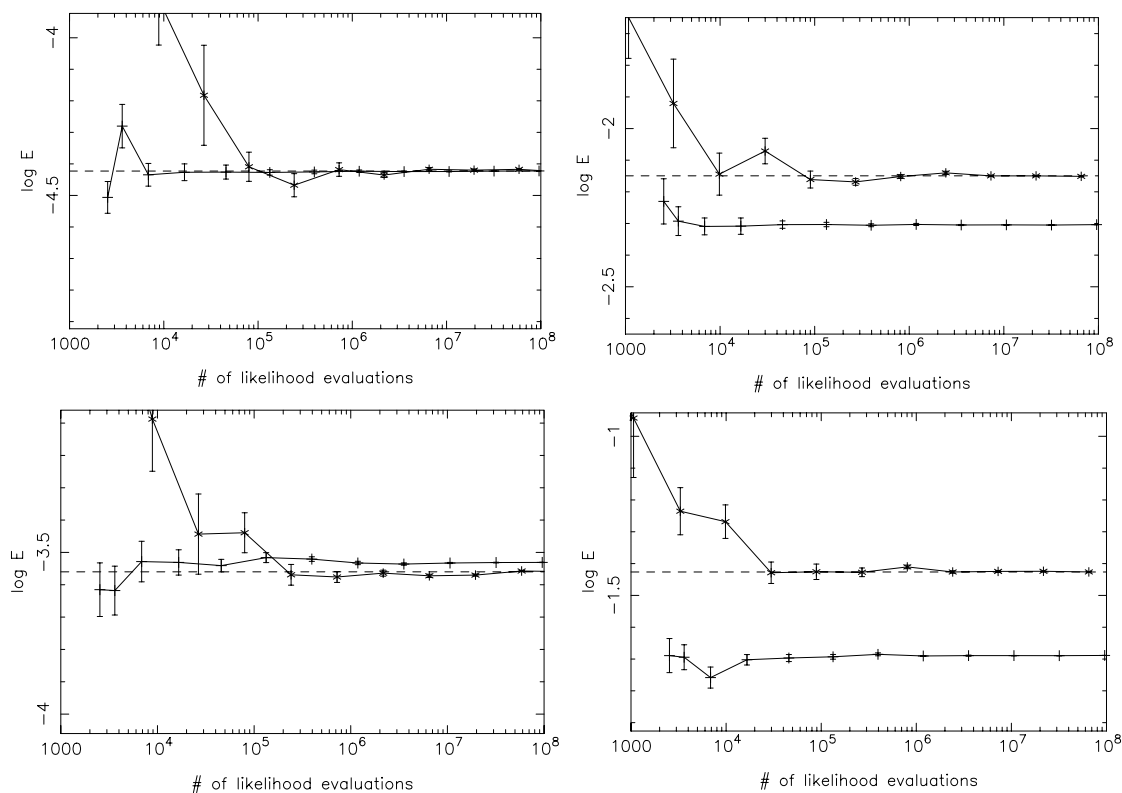

Figure 1. This figure shows the calculated evidence as a function of the number of likelihood evaluations. Note that the horizontal axis is logarithmic. The star-centred line corresponds to the thermodynamic integration. The cross-centred lines are the analytical methods with (**upper panels**) and without (**lower panels**) non-Gaussian corrections applied. The horizontal dashed line is the number obtained by the direct integration. The upper two panels correspond to $\mathcal{L}_g$, while the lower two to $\mathcal{L}_{ng}$. The left-hand side panels correspond to wide flat priors of $(-7, 10)$ on both parameters, while the right-hand side to the narrow priors of $(-2, 3)$ on both parameters. The error bars correspond to the dispersion due to the number of likelihood evaluations.

5.2. A Toy Model Comparison

We now proceed by calculating the Bayesian evidence for simple toy models with five and six parameters, shown in Table 1. The purpose is to compare results with those obtained from thermodynamic integration again, but this time using a model that bears more resemblance to a typical problem encountered in cosmology.

Table 1. The parameters used in the analytical evaluation of the toy model evidences, with five and six parameters, respectively. The maximum likelihood of the toy models is taken (arbitrarily) to be $\mathcal{L}_{\max} = 1$.

Parameter	Mean	Prior Range	Model
x_1	0.022	$[0.0001, 0.044]$	toy5, toy6
x_2	0.12	$[0.001, 0.3]$	toy5, toy6
x_3	1.04	$[0.8, 1.4]$	toy5, toy6
x_4	0.1	$[0.01, 0.3]$	toy5, toy6
x_5	3.1	$[2.6, 3.6]$	toy5, toy6
x_6	0.98	$[0.5, 1.5]$	toy6

Beginning with the five-parameter model, we first assume that it has an uncorrelated multivariate Gaussian likelihood distribution. In this case the aim is to test the thermodynamic integration method, which gives $\ln E_{\text{toy5}}^{\text{num}} = -8.65 \pm 0.03$, while the exact expression gives $\ln E_{\text{toy5}}^{\text{ana}} = -8.66$. Therefore, we conclude that the thermodynamic integration method is rather good in obtaining the correct evidence of the model. The Laplace approximation Equation (57) also fares well for uncorrelated distributions, $\ln E_{\text{toy5}}^{\text{Lap}} = -8.67$.

We now consider a likelihood function with a correlated covariance matrix C_{ij}, with the same mean values and dispersions as the previous case, but with significant correlations. The analytic formula needed, Equation (54), is no longer exact,[3] and gives $\ln E_{\text{toy5c}}^{\text{ana}} = -7.32$. For comparison thermodynamic integration gives $\ln E_{\text{toy5c}}^{\text{num}} = -7.28 \pm 0.06$, again in perfect agreement within errors. In this case the Laplace approximation fails significantly, $\ln E_{\text{toy5c}}^{\text{Lap}} = -6.89$, the reason being that the correlations chosen bring the posterior into significant contact with the edges of the priors.

Let us now return to the uncorrelated case and include a new parameter, x_6, as in Table 1, and evaluate the different evidences that appear because of this new parameter, in order to see the sensitivity to systematic errors in the evaluation of the Bayesian evidence and their effects on model comparison. The numerical result is $\ln E_{\text{toy6}}^{\text{num}} = -10.75 \pm 0.03$, while the exact analytical expression gives $\ln E_{\text{toy6}}^{\text{ana}} = -10.74$, in perfect agreement within errors. The Laplace approximation Equation (57) again fares well for uncorrelated distributions, $\ln E_{\text{toy6}}^{\text{Lap}} = -10.74$.

When the likelihood function has large correlations, and the priors are not too large, the naive Laplace approximation, Equation (57), fares less well than the analytical approximation, Equation (54).

5.3. A Real Model Comparison

In this subsection we will make use of the results obtained in ref. [12], where we evaluated the evidence for 5- and 6-parameter adiabatic models, and for three 10-parameter mixed adiabatic plus isocurvature models. The prior ranges used are given in Table 2. The latter models give a marginally better fit to the data but require more parameters, which is exactly the situation where model selection techniques are needed to draw robust conclusions. In ref. [12] we used thermodynamic integration to compute the evidence and showed that the isocurvature models were less favoured than the adiabatic ones, but only at a mild significance level.[4]

Table 2. The parameters used in the models; see ref. [12] for nomenclature and other details. For the AD-HZ model n_s was fixed to 1 and n_{iso}, δ_{cor}, α and β were fixed to 0. In the AD-n_s model, n_s also varies. Every isocurvature model holds the same priors for the whole set of parameters.

Parameter	Mean	Prior Range	Model
ω_b	0.022	$[0.018, 0.032]$	AD-HZ,AD-n_s,ISO
ω_{dm}	0.12	$[0.04, 0.16]$	AD-HZ,AD-n_s,ISO
θ	1.04	$[0.98, 1.10]$	AD-HZ,AD-n_s,ISO
τ	0.17	$[0, 0.5]$	AD-HZ,AD-n_s,ISO
$\ln[10^{10}\mathcal{R}_{rad}]$	3.1	$[2.6, 4.2]$	AD-HZ,AD-n_s,ISO
n_s	1.0	$[0.8, 1.2]$	AD-n_s,ISO
n_{iso}	1.5	$[0, 3]$	ISO
δ_{cor}	1.5	$[-0.14, 0.4]$	ISO
$\sqrt{\alpha}$	0	$[-1, 1]$	ISO
β	0	$[-1, 1]$	ISO

Beginning with the simplest adiabatic model, which uses the Harrison–Zel'dovich spectrum, we have used the analytical formulae above, Equation (54), together with the covariance matrix provided by the cosmoMC programme [21], and obtained $\ln E_{ad}^{ana} = -854.07$, while the thermodynamic integration gave $\ln E_{ad}^{num} = -854.1 \pm 0.1$ [12]. The agreement is excellent; this is because the distribution function for the adiabatic model is rather well-approximated by a Gaussian, and the priors are rather large, so the formula Equation (54) is very close to that obtained in the Laplace approximation, $\ln E_{ad}^{Lap} = -854.08$.

However the analytic method fares less well for the adiabatic model with varying n_s, with both the analytical and Laplace methods giving $\ln E_{AD-n_s} = -853.4$, while the numerical method gives the smaller value -854.1, a discrepancy of near unity.

Turning now to the isocurvature cases, we found an extremely good result for the CDI model, gaining from Equation (54) the value $\ln E_{cdi}^{ana} = -855.08$, while the thermodynamic integration gives $\ln E_{cdi}^{num} = -855.1 \pm 0.1$. This is surprising, given the relatively large non-Gaussianities for at least three variables: n_{iso}, β and δ_{cor}, whose priors are not centred with respect to the mean. However the NID case shows much less agreement, with a discrepancy of 0.6. This suggests that the closeness of the CDI comparison is to some extent a statistical fluke, with the underlying method less accurate.

A summary of the different models can be found in Table 3.

Table 3. The different models, both toy and real, with their maximum likelihoods and evidences.

Model	$\ln \mathcal{L}^{max}$	$\ln E^{num}$	$\ln E^{ana}$	$\ln E^{Lap}$
toy5	0	-8.65 ± 0.03	-8.66	-8.67
toy5c	0	-7.28 ± 0.06	-7.32	-6.89
toy6	0	-10.75 ± 0.03	-10.74	-10.74
toy6c	0	-9.73 ± 0.06	-9.71	-9.63
AD	-840.78	-854.1 ± 0.1	-854.1	-854.1
AD-n_s	-838.50	-854.1 ± 0.1	-853.4	-853.4
CDI	-838.05	-855.1 ± 0.2	-855.1	-854.5
NID	-836.60	-855.1 ± 0.2	-854.5	-854.5
NIV	-842.53	-855.1 ± 0.3	-854.9	-854.9

5.4. Savage–Dickey Method

Another numerical method for evidence calculation is the Savage–Dickey method, first described in ref. [22] and recently used in ref. [20]. This technique allows one to calculate the evidence ratio of two models from a simple and quick analysis of the Markov chains used for parameter estimation, provided that the models are nested; i.e., that one of them is included in the parameter space of the other. For instance, the AD model is nested within the AD-n_s model, and the AD and AD-n_s models are both nested within the CDI, NID and NIV ones.

In the context of Markov chains, the Savage–Dickey method is essentially a measure of how much time the sampler spends in the nested model, weighted by the respective volumes of the two models. When the outer model has extra parameters, this method relies on approximating the nested model as a model with negligibly narrow priors in directions of extra parameters. We note, however, that when many extra parameters are present, this method must fail for reasons similar to why those with grid-based parameter estimation approaches fail with models with many parameters. The MCMC parameter estimation simply does not have high enough dynamic range to probe the two models given the large prior volume ratio.

The AD and AD-n_s models differ by one parameter. Using the same AD+n_s samples as for the analytical method (i.e., the samples from which we extracted the covariance matrix), we obtained $\ln(E_{AD}/E_{AD+n_s}) = 0.03$. The result from the precise thermodynamic integration, $\ln(E_{AD}/E_{AD-n_s}) = 0 \pm 0.1$ is in excellent agreement. The AD-n_s and CDI (or NID, NIV) models differ by four parameters. With most simple choices of parametrization (including in particular the isocurvature and cross-correlation tilts), the AD-n_s is not a point, but a hyper-surface within the parameter space of the isocurvature models (i.e., $\alpha = 0$ and the other three parameters act as dummy, unconstrained, parameters which do not affect the evidence). In these cases, the evidence ratios given by the Savage–Dickey method do not converge as the priors of the extra parameters are tightened up around the nested model, although they match thermodynamically determined values to within a unit of $\ln E$.

6. Discussion and Conclusions

We have developed an analytical formalism for computing the Bayesian evidence in the case of an arbitrary likelihood distribution with a hierarchy of non-Gaussian corrections, and with arbitrary top-hat priors, centred or uncentred. This analysis can be of great help for the problem of model comparison in the present context of cosmology where observational data is still unable to rule out most extensions of the standard model based on the ΛCDM inflationary paradigm.

As an application of the exact and approximate formulae obtained for the Bayesian evidence of a model with approximately Gaussian likelihood distributions, we have compared the value predicted analytically with that computed with a time-consuming algorithm based on the thermodynamic integration approach. The values analytically obtained agree surprisingly well with those obtained numerically. While one can estimate the magnitude of the higher-order corrections for the analytical formulae, it is very difficult to estimate the systematic effects of the numerical approach. Thus, with this analytical method we can test for systematics in the thermodynamic integration approach. So far, the values obtained agree, so it seems that the numerical approach is a good tool for estimating the evidence. However, it takes considerable effort and machine time to do the correct evaluation, and therefore we propose the use of the analytical estimate, whose corrections are well under control, in the sense that one can compute the next order corrections and show that they are small.

Data Availability Statement: There is no data associated with this work.

Conflicts of Interest: The author declares no conflict of interest.

Notes

[1]   An extension to Gaussian priors should be feasible, but not one to arbitrary priors.

[2]   Note that, for scalar quantities, Einstein notation for the sum over free indices is assumed.

[3]   One could rotate the parameter basis to remove the correlations, but then the priors would not be top-hats.

[4]   Recently, Trotta [20] used a different technique to analyse a restricted class of isocurvature model featuring just one extra parameter, and found it highly disfavoured. The different conclusion is primarily due to the very different prior he chose on the isocurvature amplitude, such that almost all the models under the prior are dominated by isocurvature models and in poor agreement with the data.

References

1. Jeffreys, H. *Theory of Probability*, 3rd ed.; Oxford University Press: Oxford, UK, 1961.
2. MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK 2003.
3. Jaffe, A. H0 and odds on cosmology. *Astrophys. J.* **1996**, *471*, 24.
4. Drell, P.S.; Loredo, T.J.; Wasserman, I. Type Ia supernovae, evolution, and the cosmological constant. *Astrophys. J.* **2000**, *530*, 593. [CrossRef]
5. John, M.V.; Narlikar, J.V. Comparison of cosmological models using Bayesian theory. *Phys. Rev. D* **2002**, *65*, 043506. [CrossRef]
6. Hobson, M.P.; Bridle, S.L.; Lahav, O. Combining cosmological data sets: Hyperparameters and Bayesian evidence. *Mon. Not. R. Astron. Soc.* **2002**, *335*, 377. [CrossRef]
7. Slosar, A.; Carreira, P.; Cleary, K.; Davies, R.D..; Davis, R.J.; Dickinson, C.; Genova-Santos, R.; Grainge, K.; Gutierrez, C.M.; Hafez, Y.A.; et al. Cosmological parameter estimation and Bayesian model comparison using Very Small Array data. *Mon. Not. R. Astron. Soc.* **2003**, *341*, L29. [CrossRef]
8. Saini, T.D.; Weller, J.; Bridle, S.L. Revealing the nature of dark energy using Bayesian evidence. *Mon. Not. R. Astron. Soc.* **2004**, *348*, 603. [CrossRef]
9. Niarchou, A.; Jaffe, A.H.; Pogosian, L. Large-scale power in the CMB and new physics: An analysis using Bayesian model comparison. *Phys. Rev. D* **2004**, *69*, 063515. [CrossRef]
10. Marshall, P.; Rajguru, N.; Slosar, A. Bayesian evidence as a tool for comparing datasets. *Phys. Rev. D* **2006**, *73*, 067302. [CrossRef]
11. Liddle, A.R. How many cosmological parameters? *Mon. Not. R. Astron. Soc.* **2004**, *351*, L49–L53. [CrossRef]
12. Beltrán, M.; García-Bellido, J.; Lesgourgues, J.; Liddle, A.R.; Slosar, A. Bayesian model selection and isocurvature perturbations. *Phys. Rev. D* **2005**, *71*, 063532. [CrossRef]
13. Ó'Ruanaidh, J.J.K.; Fitzgerald, W.J. *Numerical Bayesian Methods Applied to Signal Processing*; Springer: New York, NY, USA, 1996.
14. Hobson, M.P.; McLachlan, C. A Bayesian approach to discrete object detection in astronomical data sets. *Mon. Not. R. Astron. Soc.* **2003**, *338*, 765. [CrossRef]
15. Skilling, J. Nested sampling. *AIP Conf. Proc.* **2004**, *735*, 395.
16. Handley, W.J.; Hobson, M.P.; Lasenby, A.N. POLYCHORD: Nested sampling for cosmology. *Mon. Not. R. Astron. Soc.* **2015**, *450*, L61. [CrossRef]
17. Xie, W.; Lewis, P.O.; Fan, Y.; Kuo, L.; Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **2011**, *60*, 150. [CrossRef] [PubMed]
18. Maturana-Russel, P.; Meyer, R.; Veitch, J.; Christensen, N. Search for the isotropic stochastic background using data from Advanced LIGO's second observing run. *Phys. Rev. D* **2019**, *99*, 084006.
19. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773. [CrossRef]
20. Trotta, R. Applications of Bayesian model selection to cosmological parameters. *Mon. Not. R. Astron. Soc.* **2007**, *378*, 72. [CrossRef]
21. Lewis, A.; Bridle, S. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Phys. Rev. D* **2002**, *66*, 103511. [CrossRef]
22. Dickey, J.M. The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. *Ann. Math. Stat.* **1971**, *42*, 204. [CrossRef]