

The look-elsewhere effect from a unified Bayesian and frequentist perspective

To cite this article: Adrian E. Bayer and Uroš Seljak JCAP10(2020)009

View the [article online](#) for updates and enhancements.

Recent citations

- [Matched filtering with non-Gaussian noise for planet transit detections](#)
Jakob Robnik *et al*
- [On the absence of backplash analogues to NGC 3109 in the CDM framework](#)
Indranil Banik *et al*
- [Efficient Likelihood Ratio Confidence Intervals using Constrained Optimization](#)
Gregor Reich and Kenneth L. Judd



IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

The look-elsewhere effect from a unified Bayesian and frequentist perspective

Adrian E. Bayer^a and Uroš Seljak^{a,b}

^aBerkeley Center for Cosmological Physics, Department of Physics,
University of California, Berkeley,
Berkeley, CA 94720, U.S.A.

^bPhysics Division, Lawrence Berkeley National Laboratory,
1 Cyclotron Road, Berkeley, CA 94720, U.S.A.

E-mail: abayer@berkeley.edu, useljak@berkeley.edu

Received July 29, 2020

Revised August 20, 2020

Accepted August 25, 2020

Published October 2, 2020

Abstract. When searching over a large parameter space for anomalies such as events, peaks, objects, or particles, there is a large probability that spurious signals with seemingly high significance will be found. This is known as the look-elsewhere effect and is prevalent throughout cosmology, (astro)particle physics, and beyond. To avoid making false claims of detection, one must account for this effect when assigning the statistical significance of an anomaly. This is typically accomplished by considering the trials factor, which is generally computed numerically via potentially expensive simulations. In this paper we develop a continuous generalization of the Bonferroni and Šidák corrections by applying the Laplace approximation to evaluate the Bayes factor, and in turn relating the trials factor to the prior-to-posterior volume ratio. We use this to define a test statistic whose frequentist properties have a simple interpretation in terms of the global p -value, or statistical significance. We apply this method to various physics-based examples and show it to work well for the full range of p -values, i.e. in both the asymptotic and non-asymptotic regimes. We also show that this method naturally accounts for other model complexities such as additional degrees of freedom, generalizing Wilks' theorem. This provides a fast way to quantify statistical significance in light of the look-elsewhere effect, without resorting to expensive simulations.

Keywords: dark matter experiments, gravitational waves / experiments, non-gaussianity, particle acceleration

ArXiv ePrint: [2007.13821](https://arxiv.org/abs/2007.13821)

Contents

1	Introduction	1
2	Bayesian posterior inference and hypothesis testing	3
2.1	Maximum a Posteriori (MAP) estimation	4
2.2	Maximum Posterior Mass (MPM) estimation	4
2.3	Hypothesis testing with MPM	5
3	From Bayesian to frequentist hypothesis testing	7
3.1	Maximum Posterior Significance (MPS) estimation	8
3.2	Multiple degrees of freedom	12
4	Example I: resonance searches	13
5	Example II: white noise	17
6	Example III: non-Gaussian models of cosmological inflation	19
7	Conclusions	21
A	Derivation of the CCDF of \hat{q}_S	22

1 Introduction

A common problem in statistical analysis is to find evidence for a physical signal in a large, continuous parameter space, where the true position of the signal is not known a priori. By searching over a wide parameter space one increases the probability of finding large signals caused by random statistical fluctuations, as opposed to a physical source. This is known as the look-elsewhere effect — or sometimes the “problem of multiple comparisons” in discrete cases — and must be accounted for when performing a hypothesis test [1, 2]. Ignoring this effect would lead to an overestimation of the statistical significance, sometimes by a considerable amount, and thus incorrectly concluding the detection of a physical signal.

The look-elsewhere effect is prominent throughout (astro)particle physics and cosmology. One of the most commonly known occurrences is in collider searches for new particles, for example it was a key consideration in the Higgs boson discovery [3, 4]. In this example, one searches a large range of masses for a resonance, without a priori knowledge of the true mass of the particle. Similarly, in astrophysical searches for particles one seeks resonances in the energy flux of various astrophysical spectra, where the true energy signature of the particle is unknown. Examples include: constraining the dark matter self-annihilation cross-section via gamma ray emission from galaxy clusters [5], searching for WIMPs via charged cosmic rays [6], searching for non-baryonic dark matter via X-ray emission from the Milky Way [7], and explaining the source of high energy astrophysical neutrinos [8, 9]. In terms of cosmology, the look-elsewhere effect occurs in searches for gravitational wave signals from black hole or neutron star mergers [10–12]. Here one searches large time series for a signal,

where the time and shape of the event are unknown. A further cosmological example is searching for signatures of inflation in the primordial power spectrum [13–15].

The look-elsewhere effect is also prevalent in other areas of physics and beyond, for example: in astronomy it occurs when detecting exoplanets via stellar photometry, where the period and phase of the planets’ transits are unknown (e.g. [16]); in biology it occurs when considering large DNA sequences to study genetic association [17, 18]; in medicine it occurs when testing the effectiveness of drugs in clinical trials [19]; and in theology it occurs when attempting to find hidden prophecies in religious texts [20]. Therefore, given the apparent ubiquity of the look-elsewhere effect, there is much motivation for a fast method to account for it.

Many simple general methods exist to mitigate for the look-elsewhere effect in the case of discrete problems, for example if one is testing multiple drugs for their effectiveness at treating a disease [19]. The number of drugs tested, more generally known as the trials factor, quantifies the extent of the look-elsewhere effect. The larger the trials factor, i.e. the more drugs tested, the larger the chance of a false positive arising due to a statistical fluctuation. Methods such as the Bonferroni correction [21] and Šidák correction [22] use the trials factor to correct the conclusions of a hypothesis test in light of this effect. There is however no unique definition of the trials factor when searching a continuous parameter space for a signal, making it unclear how to implement these corrections in such cases. Therefore, a common, brute-force approach to account for the look-elsewhere effect in continuous problems is to perform many simulations of an experiment assuming there is no signal. One can then estimate the p -value of a chosen test statistic, usually related to the maximum likelihood, and in turn define a relation between the significance of a signal and the test statistic. This means that to conclude a detection at the 5-sigma level, corresponding to a p -value of order 10^{-7} , one would need to simulate more than $\sim 10^7$ realizations of the experimental data, which is computationally expensive. A faster method, developed in the context of high energy physics, is to approximate the asymptotic form of the p -value by counting upcrossings, requiring fewer simulations [23]. In both of these cases new simulations are required each time a new model is considered, and the simulations may not be an accurate representation of the data. In this paper we seek an approach that can be directly applied to experimental data, without the need for simulations.

Our approach applies Bayesian logic to tackle the look-elsewhere effect. The Bayesian evidence is equal to the prior-weighted average of the likelihood over the parameter space, which can be considerably lower than the maximum likelihood if the prior is broad. This integration over the prior accounts for the look-elsewhere effect by penalizing large prior volumes. When considering large prior volumes, the likelihood is typically multimodal, with most of the peaks corresponding to noise fluctuations rather than physical sources. In order to estimate the location of a physical signal, and its associated statistical significance, one typically considers a point estimator, such as the maximum a posteriori (MAP) estimator which maximizes the posterior density. By applying the Laplace approximation, we introduce a Bayesian generalization of the MAP estimator, referred to as the maximum posterior mass (MPM) estimator, which corrects the MAP estimator by the prior-to-posterior volume ratio. Then, by drawing an analogy between Bayesian and frequentist methodology, we present a hybrid of the MAP and MPM estimators, called the maximum posterior significance (MPS) estimator, which determines the most significant peak in light of the look-elsewhere effect. The frequentist properties of the MPS estimator are shown to be independent of the look-elsewhere effect, providing a universal way to quantify the p -value, or statistical significance, without the need for expensive simulations.

The outline of this paper is as follows. In section 2 we review Bayesian posterior inference and hypothesis testing for a multimodal posterior, by discussing MAP estimation and then introducing MPM estimation. We then draw an analogy between Bayesian and frequentist philosophy in section 3 to motivate MPS estimation as the appropriate technique to tackle the look-elsewhere effect. The following three sections then apply this method to various examples: section 4 considers a resonance search, which can be thought of as a toy example of a collider or astrophysical particle search; section 5 considers a white noise time series, which can be thought of as a toy example of a gravitational wave search; and section 6 considers a search for non-Gaussian models of cosmological inflation using Planck data [24]. Note that section 4 is the main example, as it illustrates the key advantages of MPS, with the other examples complementary. Finally, we summarize and conclude in section 7.

2 Bayesian posterior inference and hypothesis testing

Two of the main tasks of Bayesian statistical analysis are posterior inference and hypothesis testing. Consider a model with parameters $\mathbf{z} = \{z_j\}_{j=1}^M$, and data $\mathbf{x} = \{x_i\}_{i=1}^{N_d}$ that depends on \mathbf{z} . The inference of \mathbf{z} is given by its posterior

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}, \quad (2.1)$$

where $p(\mathbf{x}|\mathbf{z})$ is the likelihood of the data, $p(\mathbf{z})$ is the prior of \mathbf{z} , and $p(\mathbf{x}) = \int d\mathbf{z} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ is the Bayesian evidence, also known as the normalization, marginal likelihood, or partition function.¹ Typically, one can evaluate the joint probability $p(\mathbf{x}, \mathbf{z})$, but not the evidence, which makes the posterior inference analytically intractable. This is usually handled using simple approximations or Monte Carlo Markov Chain methods [25].

A related problem is that of a hypothesis testing. In this case there are two different hypotheses, H and H_0 , each with their own, potentially different, set of model parameters. In Bayesian methodology, hypothesis testing is performed using the Bayesian evidence ratio of the two hypotheses, which gives the Bayes factor

$$B \equiv \frac{p(\mathbf{x}|H)}{p(\mathbf{x}|H_0)}, \quad (2.2)$$

where the Bayesian evidence, or marginal likelihood, for hypothesis H is given by

$$p(\mathbf{x}|H) = \int d\mathbf{z} p(\mathbf{x}|\mathbf{z}, H)p(\mathbf{z}|H). \quad (2.3)$$

If the prior on each hypothesis is equal, i.e. $p(H) = p(H_0) = 0.5$, then the Bayes factor is equal to the posterior odds ratio, $B = p(H|\mathbf{x})/p(H_0|\mathbf{x})$.

The Bayesian evidence and Bayes factor are also analytically intractable and harder to evaluate than posteriors, especially for high dimensional \mathbf{z} , although recent numerical methods such as Gaussianized Bridge Sampling [26] have made the problem easier. For the sake of exposition we will not consider such methods in this work, but instead use analytical approximations that give the Bayes factor an intuitive meaning. It is worth keeping in mind however that the full Bayes factor calculation can always be performed numerically, without any approximations.

¹Note that throughout this paper we use the letter p to refer to likelihood, prior, posterior, and evidence, even though they each refer to different functions. This notation has the advantage of making the probabilistic nature of each function clear, while the identity of the function should be clear from its argument.

2.1 Maximum a Posteriori (MAP) estimation

Given the analytical intractability of posterior inference and hypothesis testing, one often chooses an estimator to extract useful information from the posterior. A common estimator is the maximum a posteriori (MAP) point estimator, which corresponds to the global maximum of the posterior. If the prior is flat, as it will always be in this paper, this equals the maximum likelihood estimator (MLE), which maximizes the likelihood. Mathematically, MAP is defined via

$$\text{MAP} : \arg \max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}), \quad (2.4)$$

i.e. the MAP is located at the argument, \mathbf{z} , that maximizes the posterior. For the purpose of comparing data to a null hypothesis, a useful quantity to define is

$$q_L(\mathbf{z}) \equiv 2 \ln \frac{p(\mathbf{x}|\mathbf{z})}{p(\mathbf{x}|\mathbf{z}_n)}, \quad (2.5)$$

where \mathbf{z}_n represents the values of the parameters under the null hypothesis, and a subscript of L is used because the argument of the logarithm is the Likelihood ratio. To assess the significance of a result one considers the maximum value of q_L , which in the case of a flat prior is equal to q_L evaluated at the MAP: $\hat{q}_L = q_L(\mathbf{z}_{\text{MAP}})$. For a Gaussian likelihood, this is equal to the chi-squared (χ^2), and in the absence of the look-elsewhere effect $\sqrt{\hat{q}_L}$ typically gives the statistical significance. However, we will see that this test statistic greatly suffers from the look-elsewhere effect. This is because q_L effectively assumes the prior on \mathbf{z} to be a delta function located at the best fit value. Such a prior is unreasonable because it has been determined a posteriori (after seeing the data), and ignores the fact that when looking elsewhere in parameter space the prior will be broad and in turn penalize the significance. We will now explore this in more depth, and later discuss what a reasonable prior means.

2.2 Maximum Posterior Mass (MPM) estimation

MAP is often a good point estimator in low dimensions if there is a single mode in the posterior. However, if the posterior has several modes, a more reasonable point estimator associates with the highest posterior mass. We refer to this as the maximum posterior mass (MPM) estimator.

For the purposes of this work, we will consider the example of a multimodal posterior consisting of a sum of multivariate Gaussian distributions; this has been shown to be a good approximation in many practical cases [27]. We thus consider a posterior of the following Gaussian mixture form,

$$p(\mathbf{z}|\mathbf{x}) = \sum_l w^l N(\mathbf{z}; \boldsymbol{\mu}^l, \boldsymbol{\Sigma}^l), \quad (2.6)$$

where $N(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Note that working with a posterior of this form is equivalent to applying the Laplace approximation to a general multimodal posterior in the upcoming derivations. In this model, the mass of mode l is proportional to the weight w^l , which is normalized such that $\sum_l w^l = 1$.

Assuming that only one component contributes at each peak, the weight of mode l is given by evaluating the posterior at the location of the mode, $\mathbf{z} = \boldsymbol{\mu}^l$,

$$\ln w^l = \ln p(\boldsymbol{\mu}^l|\mathbf{x}) - \ln N(\boldsymbol{\mu}^l; \boldsymbol{\mu}^l, \boldsymbol{\Sigma}^l) = \ln p(\boldsymbol{\mu}^l|\mathbf{x}) + \frac{1}{2} \left[\ln \det \boldsymbol{\Sigma}^l + M \ln(2\pi) \right]. \quad (2.7)$$

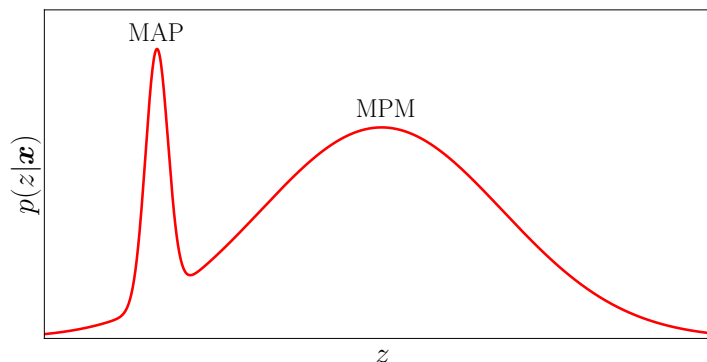


Figure 1. Plot of a bimodal Gaussian posterior for a 1d example in which 90% of the posterior mass is assigned to the right peak and 10% to the left. MPM yields the mode that maximizes the posterior mass and is close to the true mean, whereas MAP maximizes the posterior density and can be distant from the mean and represent only a small fraction of posterior mass.

Given the aforementioned analytical intractability of the posterior, we multiply each weight by the normalization $p(\mathbf{x})$ to give a quantity that can be readily computed. We call this quantity the mass m^l , and it is defined by

$$\ln m^l \equiv \ln w^l + \ln p(\mathbf{x}) = \ln p(\mathbf{x}|\boldsymbol{\mu}^l) + \ln p(\boldsymbol{\mu}^l) + \frac{1}{2} \left[\ln \det \boldsymbol{\Sigma}^l + M \ln(2\pi) \right]. \quad (2.8)$$

Thus the mass of each mode is equal to the likelihood multiplied by the product of the prior density and the posterior volume at the peak, where the posterior volume is defined as

$$V_{\text{posterior}} \equiv (2\pi)^{M/2} \sqrt{\det \boldsymbol{\Sigma}}. \quad (2.9)$$

The MPM estimator corresponds to the mode with the highest mass, thus

$$\text{MPM} : \arg \max_{\mathbf{z}} \left[p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \sqrt{\det \boldsymbol{\Sigma}(\mathbf{z})} \right]. \quad (2.10)$$

To determine the MPM mode one would compute the $\ln m^l$ by first finding the positions of all local posterior maxima $\boldsymbol{\mu}^l$, and then computing $\boldsymbol{\Sigma}^l$ using the inverse of the Hessian at each peak. Qualitatively, MPM corresponds to maximizing the posterior density multiplied by the posterior volume $\sim \sqrt{\det \boldsymbol{\Sigma}}$, whereas MAP only maximizes the former. It is apparent that if there are multiple modes in the posterior, the one that has the largest posterior mass does not necessarily have the largest posterior density, as shown in figure 1. In some situations the MPM mode will dominate the posterior mass such that the MPM mode alone gives a useful way to summarize the posterior.

2.3 Hypothesis testing with MPM

Consider a model with parameters z_1, z_2, \dots, z_M , with z_1 corresponding to the amplitude of a feature, and z_2, \dots, z_M corresponding to the properties of the feature. For example, z_1 might correspond to the amplitude of a signal detected in a time series at time z_2 . We will use the notation $\mathbf{z} \equiv (z_1, \dots, z_M)^T$ as the vector of all parameters, and $\mathbf{z}_{>1} \equiv (z_2, \dots, z_M)^T$ as the subvector of non-amplitude parameters, i.e. excluding z_1 . A typical analysis would scan over $\mathbf{z}_{>1}$, finding the best fit value for the amplitude z_1 at each point, giving rise to a multimodal posterior.

In this work we wish to determine whether or not a dataset contains a true anomaly. In the language of hypothesis testing, we wish to compare the hypothesis that there is an anomaly H , corresponding to $z_1 > 0$, to the null hypothesis that there is no anomaly H_0 , corresponding to $z_1 = 0$. We assume the common case that the parameters of H_0 are a subset of the parameters of H , with H reducing to H_0 when $z_1 = 0$. There may also be parameters other than \mathbf{z} that are common to both models, but these are of secondary importance when considering the look-elsewhere effect and we drop these from the notation.

Using equation (2.8) with $\sum w^l = 1$ implies that the Bayesian evidence for hypothesis H is given by

$$p(\mathbf{x}|H) = \int d\mathbf{z} p(\mathbf{x}|\mathbf{z}, H)p(\mathbf{z}|H) = \sum_l m^l, \quad (2.11)$$

where the m^l correspond to the masses under hypothesis H . Hence, each mode contributes its mass to the evidence. It follows that the mass of mode l corresponds to the Laplace approximation of the evidence integral in equation (2.11), integrated over the region of the mode. Because the null hypothesis does not depend on $\mathbf{z}_{>1}$, the evidence for the null hypothesis is given by the likelihood evaluated at $z_1 = 0$, that is $p(\mathbf{x}|H_0) = p(\mathbf{x}|z_1 = 0) \equiv p_0(\mathbf{x})$. Together with equation (2.11) this gives the Bayes factor

$$B \equiv \frac{p(\mathbf{x}|H)}{p(\mathbf{x}|H_0)} = \frac{1}{p_0(\mathbf{x})} \sum_l m^l \equiv \sum_l b^l, \quad (2.12)$$

where b^l is defined as the contribution of mode l to the Bayes factor. Using equation (2.8) gives

$$b^l = \frac{p(\mathbf{x}|\boldsymbol{\mu}^l)}{p_0(\mathbf{x})} p(\boldsymbol{\mu}^l) (2\pi)^{M/2} \sqrt{\det \boldsymbol{\Sigma}^l} = \frac{p(\mathbf{x}|\boldsymbol{\mu}^l)}{p_0(\mathbf{x})} \frac{V_{\text{posterior}}(\boldsymbol{\mu}^l)}{V_{\text{prior}}(\boldsymbol{\mu}^l)}, \quad (2.13)$$

where we have introduced the effective volume of the prior at $\boldsymbol{\mu}^l$ as,

$$V_{\text{prior}}^{-1}(\boldsymbol{\mu}^l) \equiv p(\boldsymbol{\mu}^l), \quad (2.14)$$

appropriate for the case of a narrow posterior relative to the prior. In the remainder of this paper we will drop the $\boldsymbol{\mu}^l$ dependence of the prior volume, as appropriate for a flat prior on \mathbf{z} .

Intuitively, one can think of each b^l as the Bayes factor one would get if mode l were the only mode in the posterior. If the maximum b^l is sufficiently large, it alone can provide a useful approximation to the Bayes factor, meaning the MPM mode dominates the Bayes factor. The first ratio on the right hand side of equation (2.13) corresponds to the likelihood ratio of the signal hypothesis to the null hypothesis, evaluated at the location of the peak, $\mathbf{z} = \boldsymbol{\mu}^l$. This is greater than or equal to 1 since adding parameters to the null hypothesis can only improve the fit. The second ratio gives the ratio of the posterior volume to the prior volume at the peak, which is always less than 1. This acts as a penalty to the likelihood ratio, often referred to as the Occam's razor penalty [28], or model complexity penalty, and compensates for the look-elsewhere effect in the case of a multimodal posterior. The higher the prior-to-posterior volume ratio, the higher the chance that peaks with a high likelihood will occur because of statistical fluctuations, thus the larger the penalty required to compensate.

Just as q_L is the estimator associated with MAP, we can define $q_b \equiv 2 \ln b$ as the estimator associated with MPM, such that

$$q_b = q_L - 2 \ln \frac{V_{\text{prior}}}{V_{\text{posterior}}}. \quad (2.15)$$

The MPM mode corresponds to the mode with maximum q_b . This illustrates how the MAP estimator ignores the look-elsewhere penalty by effectively considering the posterior and prior to be overlapping delta functions, which presumes a priori knowledge of the parameters and gives a prior-to-posterior volume ratio of unity.

An interesting question to consider is whether one can relate q_b to the look-elsewhere corrected statistical significance in a frequentist sense. In the absence of the look-elsewhere effect, the significance is given by $\sqrt{q_L}$, but simply taking $\sqrt{q_b}$ as the look-elsewhere corrected significance would not be correct. In the next section we turn to a frequentist description of the look-elsewhere effect to motivate a new estimator which applies a small modification to q_b and has a simple interpretation in terms of the significance, or p -value.

Before ending this section we discuss the choice of priors appropriate for a look-elsewhere analysis. If one has no prior knowledge regarding the location of an anomaly, then a uniform prior for the $\mathbf{z}_{>1}$ parameters is appropriate. If the prior is wide and posterior narrow this induces a large look-elsewhere effect. This choice of prior is not controversial. On the other hand, the choice of prior for the amplitude parameter z_1 is less clear. If one has no prior knowledge of the signal amplitude, then one should be open to a signal of any size, however one does not want the amplitude prior to induce a look-elsewhere penalty. In Bayesian hypothesis testing the amplitude parameter is treated analogously to the other parameters, thus if one uses too broad an amplitude prior it will induce an unwanted look-elsewhere penalty, whereas if one chooses too narrow an amplitude prior one risks discounting a large signal. Based on this we rewrite b in the following form, explicitly separating the marginalization over $\mathbf{z}_{>1}$ and z_1 ,

$$b = e^{q_L/2} \frac{V_{>1,\text{posterior}}}{V_{>1,\text{prior}}} \frac{V_{1,\text{posterior}}}{V_{1,\text{prior}}}. \quad (2.16)$$

The posterior volume terms are given by the covariance matrix, as in equation (2.9), and $V_{>1,\text{prior}}$ is given by the choice of prior on $\mathbf{z}_{>1}$. It thus remains to justify a choice of $V_{1,\text{prior}}$, which we will do by turning to a frequentist description of the look-elsewhere effect in the next section.

3 From Bayesian to frequentist hypothesis testing

Standard statistics literature states that Bayesian and frequentist hypothesis testing follow different methodologies and may give very different results. One famous illustration of this is the Jeffreys-Lindley “paradox” [29], however, there is much debate as to whether this is indeed a paradox and how relevant it is for scientific discourse (see [30] for a review in the context of high energy physics). While Bayesian statistics uses the Bayes factor for hypothesis testing, frequentist statistics uses the maximum likelihood ratio, or \hat{q}_L . One of the most important aspects of frequentist methodology is the computation of the false positive rate using the p -value, which quantifies how often a test statistic, for example \hat{q}_L , will take a specific value or larger under the assumptions of the null hypothesis. This has an intuitive interpretation as it directly relates to the false positive rate of the test statistic. On the other hand, Bayesian methodology rejects the p -value. The basis for this rejection is the likelihood

principle, which states that any inference about the parameters \mathbf{z} from the data \mathbf{x} can only be made via the likelihood $p(\mathbf{x}|\mathbf{z})$. When the likelihood principle is applied to testing a hypothesis with parameters \mathbf{z} one must use the marginal likelihood by integrating out these parameters — as in the Bayesian evidence of equation (2.3) — thus Bayesian methodology explicitly satisfies the likelihood principle. It is commonly argued that p -values violate the likelihood principle, because they rely on the frequentist properties of a distribution that go beyond the likelihood principle. However, the Bayes factor provides a less reliable tool for model comparison, as it is often interpreted in terms of arbitrary, model-independent scales [31], unlike the p -value which directly relates to the false positive rate.

We seek to elucidate how the answers of the two schools of statistics relate to one and other when it comes to the hypothesis testing. Both schools of statistics should give a similar, or at least related answer, when the question is phrased similarly. For uncertainty quantification it is often argued that the two schools do not answer the same question, since the Bayesian school treats data as fixed and varies the models, while the frequentist school varies the data at a fixed model. However, when it comes to hypothesis testing the distinction is less prominent: for example, when comparing two discrete hypotheses without any marginalizations, the answer in both cases gives the likelihood ratio as the optimal statistic (assuming equal prior for the two hypotheses). For continuous hypotheses it is often argued this is not possible. Here we will show that the two answers, the p -value and the Bayes factor, can be related with a specific choice of prior. It is important to emphasize that we are not claiming to equate the Bayesian and frequentist methodologies, but rather motivate a connection.

In this work we define the p -value as the probability under the null hypothesis, H_0 , of a random variable, Q , to be observed to have a value equal to or more extreme than the value observed, q . We thus use the notation $P(Q \geq q)$ for the p -value. To compute the p -value of a test statistic, one must consider how the test statistic is distributed under the null hypothesis. For the example of \hat{q}_L this distribution is not universal: scanning over continuous variables, as in the look-elsewhere effect, will modify this distribution in a model dependent manner. Moreover, increasing the model complexity in other ways, for example by including extra degrees of freedom, will further modify the distribution. To account for extra degrees of freedom, Wilks' theorem [32] provides the asymptotic distribution of \hat{q}_L for a hypothesis test where H has ν more degrees of freedom than H_0 . However, Wilks' theorem relies on technical conditions, such as the observed value not being at the edge of the interval, and does not consider the look-elsewhere effect. Generalization of Wilks' theorem for the look-elsewhere effect have been considered in [33, 34] and have been translated into a practical procedure in [23]. As a result, a frequentist approach consists of a series of considerations to determine the change in the distribution of \hat{q}_L due to different sources of model complexity. This is unlike the Bayesian methodology where all forms of model complexity are accounted for in the same way, as they are encoded into the Bayes factor. By connecting the two methodologies, we will present a test statistic whose distribution is universal, regardless of the model complexity and look-elsewhere effect.

3.1 Maximum Posterior Significance (MPS) estimation

We start by considering the typical case of one degree of freedom, corresponding to a single signal with amplitude z_1 and features described by $\mathbf{z}_{>1}$. We denote q_L maximized over the amplitude parameter only as $\tilde{q}_L(\mathbf{z}_{>1}) \equiv \max_{z_1} q_L(\mathbf{z})$, not to be confused with $\hat{q}_L \equiv \max_{\mathbf{z}} q_L(\mathbf{z})$ which is q_L maximized over all parameters. For a t -tailed test (where t is equal

1 or 2), Wilks' theorem gives the asymptotic p -value of \check{q}_L , at any position $z_{>1}$, as

$$P(\check{Q}_L \geq \check{q}_L) = \frac{t}{2} \tilde{F}_1(\check{q}_L) \xrightarrow{\check{q}_L \rightarrow \infty} \frac{t}{\sqrt{2\pi\check{q}_L}} e^{-\check{q}_L/2}, \quad (3.1)$$

where \tilde{F}_ν is the complementary cumulative distribution function (CCDF) of a chi-squared random variable with ν degrees of freedom. This maximization over z_1 at a fixed choice of $z_{>1}$ corresponds to the p -value in the absence of the look-elsewhere effect, referred to as the *local* p -value. Further maximizing over $z_{>1}$ introduces the look-elsewhere effect, which can be parameterized by multiplying by the trials factor N such that

$$P(\hat{Q}_L \geq \hat{q}_L) = N \frac{t}{\sqrt{2\pi\hat{q}_L}} e^{-\hat{q}_L/2}. \quad (3.2)$$

This is referred to as the *global* p -value. It is this form that encapsulates the Bonferroni correction [21] which effectively multiplies the local p -value by N to account for the look-elsewhere effect. For discrete problems the trials factor equals the number of trials performed. However, in the continuous case it is ill-defined, but it quantifies how the probability of finding a spurious peak increases as one looks elsewhere in the space spanned by $z_{>1}$. Accounting for the look-elsewhere effect thus requires an expression for the trials factor.

It follows from equation (3.2) that one can define a test statistic,

$$q_S = q_L - 2 \ln N + \ln 2\pi q_L - 2 \ln t \quad (3.3)$$

such that the global p -value tends to

$$P(\hat{Q}_S \geq \hat{q}_S) \rightarrow e^{-\hat{q}_S/2}, \quad (3.4)$$

as either $N \rightarrow \infty$ or $\hat{q}_S \rightarrow \infty$, so this also applies for $N = 1$. See appendix A for a derivation. Unlike \hat{q}_L , \hat{q}_S has a distribution that is independent of N — the look-elsewhere effect has been absorbed into the test statistic. Intuitively one can think of the $2 \ln N$ term as a penalty to q_L to correct for the look-elsewhere effect, while the $\ln 2\pi q_L$ term removes q_L dependent bias, ensuring the p -value depends on \hat{q}_S alone in the asymptotic limit. Thus to account for the look-elsewhere effect one need only compute \hat{q}_S and use this equation to compute the p -value. Because the p -value is a monotonically decreasing function of \hat{q}_S , one can think of selecting the peak with maximum q_S as selecting the peak with minimum p -value or maximum statistical significance. We refer to the mode with maximum q_S as the MPS mode, deferring an explanation for this nomenclature until the end of the subsection. The similarity of q_S to q_b from equation (2.15) suggests a connection between the frequentist and Bayesian pictures, and we now invoke this connection to find an expression for N and in turn generalize the Bonferroni correction to continuous parameters.

Heuristically, the Bayes factor describes the probability of the alternative hypothesis relative to the null, determined by the likelihood (as measured by \hat{q}_L), while the p -value averages its inverse over all values larger than \hat{q}_L and will be smaller than the likelihood. We expect that for higher \hat{q}_L the effect is larger because we are further into the tail of the distribution. There is no unique relation between the two, but one simple option is that the p -value scales as $B^{-1}/\hat{q}_L \approx \hat{b}^{-1}/\hat{q}_L$, where hats now indicate quantities associated with the MPS mode. Because we have the freedom to choose the prior on z_1 , we can *define* the relation between the Bayes factor and p -value as

$$\frac{\hat{b}^{-1}}{\hat{q}_L} \equiv P(\hat{Q}_L \geq \hat{q}_L). \quad (3.5)$$

Comparing equation (2.16) with equation (3.2) then gives

$$\frac{V_{>1,\text{prior}}}{\hat{V}_{>1,\text{posterior}}} \frac{V_{1,\text{prior}}}{\hat{V}_{1,\text{posterior}}} \frac{e^{-\hat{q}_L/2}}{\hat{q}_L} = N \frac{t}{\sqrt{2\pi\hat{q}_L}} e^{-\hat{q}_L/2}. \quad (3.6)$$

By requiring that this relation holds in the absence of the look-elsewhere effect, the trials factor can be identified as

$$N = \frac{V_{>1,\text{prior}}}{\hat{V}_{>1,\text{posterior}}}, \quad (3.7)$$

and the amplitude prior volume is given by

$$V_{1,\text{prior}} = t\sqrt{\hat{q}_L} \frac{\hat{V}_{1,\text{posterior}}}{\sqrt{2\pi}} = t\sqrt{\hat{q}_L} \hat{\sigma}_1 \approx t\hat{\mu}_1. \quad (3.8)$$

In the final steps we used $\hat{V}_{1,\text{posterior}} = \sqrt{2\pi}\hat{\sigma}_1$, where $\hat{\sigma}_1$ is the error on the amplitude parameter, $\hat{\mu}_1$, and that the signal-to-noise ratio obeys $\sqrt{\hat{q}_L} \approx \hat{\mu}_1/\hat{\sigma}_1$. Since the look-elsewhere effect leads to large \hat{q}_L , this prior volume will be larger than the posterior volume. This choice of amplitude prior volume ensures that there is no trials factor associated with the amplitude, as intuition would dictate. Substituting equations (3.7) and (3.8) into equation (3.3) yields

$$q_S = q_b + 2 \ln q_L. \quad (3.9)$$

Hence, we have effectively applied a modification to the MPM estimator to give a combination of the MPM and MAP estimators, so that the asymptotic p -value is neatly given by $e^{-\hat{q}_S/2}$. In the context of the look-elsewhere effect, the mode with maximum q_b will typically also be the mode with maximum q_L , and thus maximum q_S . However, this equivalence of MAP and MPM may not always be the case, as shown in figure 1.

A pure Bayesian might argue that equation (3.8) is not a valid prior, since it depends on the a posteriori amplitude parameter $\hat{\mu}_1$; however, this prior does have an intuitive justification. If a scientist is willing to consider a signal of any amplitude, the prior cannot be zero at $\hat{\mu}_1$, as it would not make sense to discard the signal. On the other hand, making the prior significantly broader than $\hat{\mu}_1$ implies the scientist has some additional information on the nature of the amplitude. When there is no justification for broadening the prior, the narrowest possible prior still consistent with the measured value can be more reasonable than arbitrarily fixing the size of the prior a priori. This choice of amplitude prior is simply designed to allow for a signal with any amplitude, without inducing an unwanted look-elsewhere penalty.

Note that the explicit dependence on \hat{q}_L and the marginal likelihood, via \hat{b} , in equation (3.5) is what makes the p -value inconsistent with the likelihood principle. One could instead consider equating \hat{b}^{-1} directly with the p -value, making it consistent with the likelihood principle. This would require an amplitude prior of $V_{1,\text{prior}} = t\hat{\sigma}_1^2/\hat{\mu}_1$, which we deem unreasonable as it is smaller than the posterior volume $\hat{V}_{1,\text{posterior}}$. We emphasize that the equality of \hat{b}^{-1}/\hat{q}_L to the p -value is not required for our approach to the look-elsewhere effect, but provides insight on the Bayesian-frequentist connection. At its core, our method considers the test statistic \hat{q}_S , from equation (3.3), and replaces the trials factor N with the prior-to-posterior volume of the non-amplitude parameters $\mathbf{z}_{>1}$. Intuitively one can think of the number of trials as the number of posterior volumes that fit within the prior volume, and this intuition suggests \hat{b}^{-1} scales linearly with the p -value.

Because the asymptotic p -value scales linearly with the prior volume, the non-asymptotic form of the p -value can be derived by dividing the prior volume into $K \gg 1$ regions and evaluating the p -value for each. Assuming independence between these regions, the product of the p -values for each region can be used to obtain p -value of the full volume. Further assuming that the asymptotic regime still applies, this gives

$$P(\hat{Q}_S \geq \hat{q}_S) = \lim_{K \rightarrow \infty} \left[1 - \left(1 - \frac{e^{-\hat{q}_S/2}}{K} \right)^K \right] = 1 - \exp \left(-e^{-\hat{q}_S/2} \right). \quad (3.10)$$

Just as equation (3.4) is a generalization of the Bonferroni correction, equation (3.10) is a generalization of the Šidák correction [22] to continuous variables. This expression generalizes the p -value into the non-asymptotic regime.

For $N \gg 1$ every realization will have a positive peak, hence even the one-tailed p -value will approach 1 for sufficiently low \hat{q}_L , which equation (3.10) predicts to be for $\hat{q}_S < 0$. In the absence of the look-elsewhere effect ($N = 1$) a one-tailed test should approach a p -value of 0.5, while equation (3.10) approaches 1 as $\hat{q}_S \rightarrow -\infty$. Thus, the non-asymptotic agreement breaks down for $t = 1$ and $N = 1$. On the other hand, if $t = 2$ and $N = 1$, substituting $q_S = q_L + \ln 2\pi q_L - 2 \ln 2$ into equation (3.10) gives

$$P(\hat{Q}_L \geq \hat{q}_L)_{N=1, t=2} = 1 - \exp \left[- \left(\frac{2}{\pi} \frac{1}{\hat{q}_L} \right)^{1/2} e^{-\hat{q}_L/2} \right]. \quad (3.11)$$

The term in the square brackets can be identified as the asymptotic expansion of $\tilde{F}_1(\hat{q}_L)$. We show the non-asymptotic agreement of this equation with $\tilde{F}_1(\hat{q}_L)$, the true two-tailed p -value for $N = 1$, in figure 2. This illustrates the ability of the generalized Šidák correction to produce correct non-asymptotic results, even in the absence of the look-elsewhere effect. Hence, although we have applied asymptotic approximations throughout the above calculations, we have obtained a result that is valid even in the non-asymptotic limit. Inverting equation (3.11) gives the significance, or number of sigma, S , as

$$S^2 \approx \hat{q}_S - \ln 2\pi \hat{q}_S + 2 \ln t, \quad (3.12)$$

with corrections of order $\mathcal{O}(\hat{q}_S^{-1})$. In the limit of $\hat{q}_S \rightarrow \infty$, the significance can be interpreted as $\sqrt{\hat{q}_S}$, in an analogous way to $\sqrt{\hat{q}_L}$ in the absence of the look-elsewhere effect. This motivates the name maximum posterior significance (MPS) as q_S depends on the posterior via the trials factor N , and is monotonically related to the significance S .

In summary, by considering a frequentist description of the look-elsewhere effect we introduced \hat{q}_S as a natural test statistic to use, such that the asymptotic p -value is given by $e^{-\hat{q}_S/2}$. We derived a general expression for the p -value which also applies in the non-asymptotic regime, and when there's no look-elsewhere effect. Adopting the prior of equation (3.8), we showed that one can write the p -value in terms of Bayes factor as \hat{b}^{-1}/\hat{q}_L . This intrinsically accounts for the look-elsewhere effect by identifying the trials factor as the prior-to-posterior volume ratio of $z_{>1}$ at the MPS mode. While one can compute the Bayes factor using a variety of methods, we will use the Laplace approximation to evaluate the posterior volume of each mode, as in section 2. To outline the step-by-step approach:

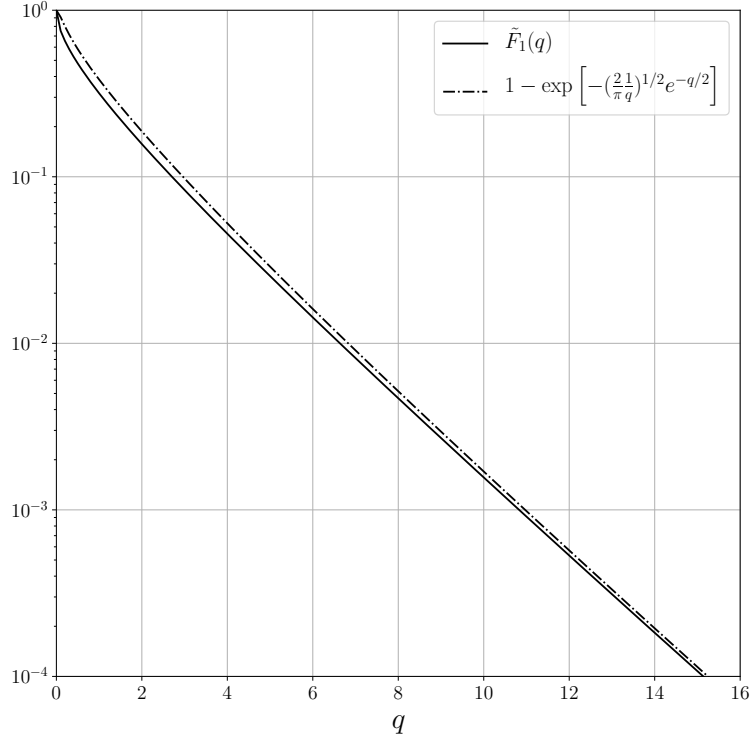


Figure 2. Equation (3.11) is a good approximation to $\tilde{F}_1(q)$ over the entire range of q . This suggests that MPS is still accurate in the absence of the look-elsewhere effect for a two-tailed test, even non-asymptotically.

Maximum Posterior Significance (MPS) estimation:

1. Scan over the space of non-amplitude parameters, $\mathbf{z}_{>1}$, locating peaks in the posterior with any amplitude, z_1 . Often only the highest few peaks are needed.
2. Compute q_L and the posterior volume, using equation (2.9), for each peak.
3. Compute q_b for each peak using equation (2.15) with the amplitude prior of equation (3.8).
4. Compute $q_S = q_b + 2 \ln q_L$ for each peak.
5. Find the peak with maximum q_S .
6. Compute the (global) p -value using equation (3.10) and significance using (3.12).

3.2 Multiple degrees of freedom

For models with multiple degrees of freedom, the frequentist approach is to apply Wilks' theorem [32]. This is valid in the asymptotic limit, where, for a two-tailed test, the local p -value is given by

$$P_\nu(\tilde{Q}_L \geq \check{q}_L) = \tilde{F}_\nu(\check{q}_L) \xrightarrow{\check{q}_L \rightarrow \infty} \frac{1}{\Gamma(\nu/2)} \left(\frac{\check{q}_L}{2} \right)^{\nu/2-1} e^{-\check{q}_L/2}, \quad (3.13)$$

for a model with ν degrees of freedom. Note that the limit assumes $q \gg \nu$, but for $\nu = 2$ it is exact for any q . Wilks' theorem can address the model complexity problem of having multiple (ν) continuous amplitude parameters. A specific example from particle physics is a decay process with ν decay channels, each with amplitude A_i ($0 \leq i \leq \nu$). In such a case $\max_{\{A_i\}} q_L(\{A_i\}, \dots) \sim \tilde{F}_\nu$. Wilks' theorem is not sufficiently general: it fails if the parameters are at the edge of their distribution, and it does not naturally handle the model complexity of the look-elsewhere effect, where one scans over a wide range of values for one or more parameters. Upon introduction of the look-elsewhere effect a frequentist would typically consider single trials distributed as $\sim \tilde{F}_\nu$, and then use a ν -dependent trials factor [23]. Thus in a frequentist approach extra degrees of freedom and the look-elsewhere effect are treated separately. On the other hand, a Bayesian approach accounts for both in the same way.

To apply the Bayesian methodology, we first reparameterize the model so that there is only a single amplitude parameter by introducing branching ratios α_i , such that each amplitude parameter is $A_i = \alpha_i z_1$, where z_1 is the total amplitude parameter and $\sum_{i=1}^\nu \alpha_i^2 = 1$. To remove the constraint we adopt rotation angles: for example, for $\nu = 2$ we can work with a phase angle ϕ , such that $\alpha_1 = \cos \phi$ and $\alpha_2 = \sin \phi$. Thus, instead of working with A_1 and A_2 and considering $\max_{A_1, A_2} q_L(A_1, A_2, \dots) \sim \tilde{F}_2$, we consider $\max_{z_1} q_L(z_1, \phi, \dots) \sim \tilde{F}_1$ with $\mathbf{z}_{>1} = (\phi, \dots)$. We can then directly apply the MPS prescription for $\nu = 1$, as in the previous subsection, by additionally marginalizing over ϕ to account for the model complexity with an additional prior-to-posterior volume penalty.

To be agnostic, one would choose a prior volume for ϕ of $V_{\phi, \text{prior}} = \pi$ (in practice a more complex prior may be appropriate, but it will typically be $\mathcal{O}(1)$). Furthermore, the average error on ϕ is typically equal to the relative error on the amplitude, thus $\sigma_\phi \approx \sigma_1/\mu_1 \approx q_L^{-1/2}$. This gives a model complexity correction of

$$\frac{V_{\phi, \text{prior}}}{\hat{V}_{\phi, \text{posterior}}} = \frac{\pi}{\sqrt{2\pi}\hat{\sigma}_\phi} = \sqrt{\pi} \left(\frac{\hat{q}_L}{2} \right)^{1/2} = \frac{\tilde{F}_2(\hat{q}_L)}{\tilde{F}_1(\hat{q}_L)}. \quad (3.14)$$

This shows that increasing the model complexity with an extra degree of freedom is accounted for in the Bayesian framework by marginalizing over ϕ . Thus, the Bayesian answer to an increase in model complexity, whether it be due to including extra degrees of freedom, or looking elsewhere, is identical: marginalization over the non-amplitude parameters $\mathbf{z}_{>1}$. The ν dependence of the local p -value in equation (3.13) can be interpreted as a Bayesian model complexity penalty: a fixed p -value corresponds to a larger \hat{q}_L as ν increases. Thus, MPS intrinsically generalizes Wilks' theorem by relating the trials factor to the prior-to-posterior volume.

4 Example I: resonance searches

To test the theory of section 3 we first consider a resonance search example. These appear in many different areas of physics, including astroparticle and high energy physics. We consider a search for a new particle whose mass and cross-section are unknown. The data \mathbf{x} could correspond to measurements of the invariant mass in the case of collider searches, or the energy flux in astroparticle searches. The probability density for a single measurement, x^i , is given by

$$p(x^i|f, x_*, \sigma_*) = f p_s(x^i|x_*, \sigma_*) + (1-f) p_b(x^i), \quad (4.1)$$

where p_s and p_b are the normalized signal and background distributions respectively, and f is the fraction of events belonging to the signal. We assume that the form of the signal

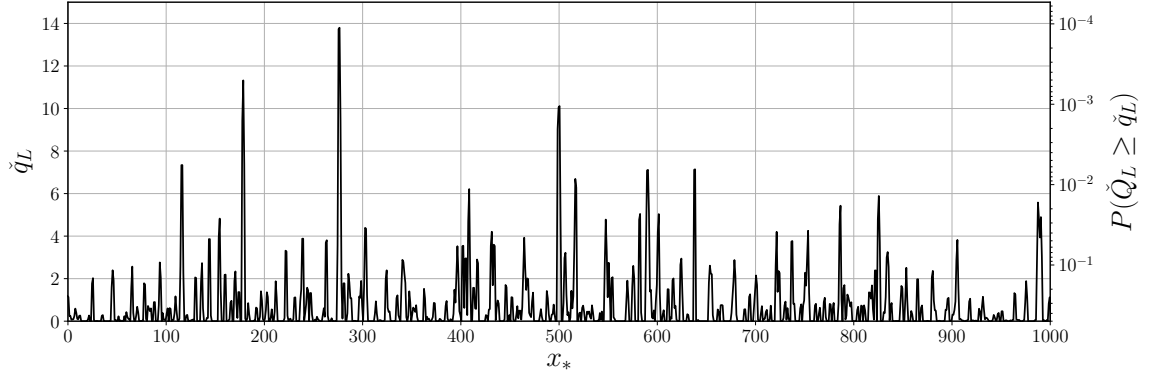


Figure 3. The local chi-squared (left axis) and local p -value (right axis) for an example data realization with true amplitude $f = 5 \times 10^{-3}$, position $x_* = 500$, and width $\sigma_* = 0.5$. While there is a peak with $\tilde{q}_L \approx 10$ at the correct position, the look-elsewhere effect leads to other, sometimes larger, peaks at random positions.

and background are known; we take the signal to be a normal distribution $p_s(x^i|x_*, \sigma_*) = N(x^i|x_*, \sigma_*)$, and the background to be a power law. Thus the resonance has position x_* and width σ_* . Given data $\mathbf{x} = \{x^i\}_{i=1}^{N_d}$, the likelihood is given by the product of the individual probability densities over the data. Using equation (4.1) this gives the likelihood as

$$p(\mathbf{x}|f, x_*, \sigma_*) = \prod_{i=1}^{N_d} [f p_s(x^i|x_*, \sigma_*) + (1-f) p_b(x^i)]. \quad (4.2)$$

Note that the Bayesian evidence under the null hypothesis is independent of the parameters, namely

$$p_0(\mathbf{x}) \equiv p(\mathbf{x}|f=0) = \prod_{i=1}^{N_d} p_b(x^i). \quad (4.3)$$

While the likelihood depends on the number of data N_d , quantities such as the p -value will have converged provided N_d is sufficiently large to resolve the resonance. Throughout this section we fix $N_d = 10V_{x_*, \text{prior}}/\sigma_*$ to ensure sufficient convergence. We note that more complex models might consider drawing N_d from a Poisson distribution, however this is unnecessary for our proof of concept.

We first consider a uniform prior on x_* , with range $(0, 10^3)$, i.e. a prior volume of $V_{x_*, \text{prior}} = 10^3$. We do not fit for σ_* and fix it to $\sigma_* = 0.5$ a priori, corresponding to the narrow-width approximation. In this case the posterior is only multimodal in the x_* dimension, thus to find peaks we split the parameter space along the x_* dimension into narrow bins of size Δx_* and compute the maximum likelihood of equation (4.2) within each bin. Ensuring Δx_* is sufficiently small, we determine the location of all peaks in the posterior, $\boldsymbol{\mu}^l$, by comparing adjacent bins. The Hessian at each peak is then computed using finite differencing, and inverted to give $\boldsymbol{\Sigma}^l$. Note, in this example we have an analytical form for the likelihood, enabling verification of the numerical computation with analytical results. The value of q_b at each peak is then computed using equation (2.15), in turn giving \hat{q}_s .

Figure 3 shows the local chi-squared and local p -value as a function of x_* for an example data realization. We use true parameters $f = 5 \times 10^{-3}$ and $x_* = 500$. Recall from equation (3.1) that the local chi-squared and p -value correspond to the values obtained by

maximizing over f at fixed x_* , i.e. they correspond to the values obtained without having corrected for the look-elsewhere effect. The local chi-squared \hat{q}_L can also be thought of as the projection of q_L onto the x_* axis. It can be seen that although there is a peak with $q_L \approx 10$ at the correct position, there are also multiple spurious peaks throughout the parameter space, with $\hat{q}_L \approx 14$ in this example. This illustrates the look-elsewhere effect: peaks with a local p -value of $\sim 10^{-4}$ are produced by noise, meaning a signal with such a local p -value should not be considered as significant as its local p -value naively suggests.

We now consider 10^5 different data realizations without a signal ($f = 0$) to study the distributions of \hat{q}_L and \hat{q}_S under the null hypothesis. The plots in figure 4 show the global p -value in terms of \hat{q}_L and \hat{q}_S for a variety of scenarios. One can think of the vertical axes as corresponding to the false positive rate (FPR) of a hypothesis test using threshold q .

We first compare three different prior volumes on x_* , $V_{x_*, \text{prior}} = 10^3, 10^2, 10^1$, to show the effectiveness of our method for large and small N . The top left plot of figure 4 shows that the p -value of \hat{q}_L has a considerable prior volume dependence. This is the look-elsewhere effect: a larger prior volume leads to a larger trials factor and thus an increased probability of finding a higher maximum likelihood. On the other hand we see that \hat{q}_S shows no prior dependence and is in good agreement with equation (3.10), even in the non-asymptotic regime.

We also investigate the variation of the p -value with the value of the width of the signal σ_* . This is shown in the top right plot of figure 4 where we consider $\sigma_* = 0.1, 0.5, 1.0$. Smaller σ_* leads to a smaller posterior volume and thus a larger trials factor. Much like the discussion above for prior volume variation, \hat{q}_L has a large σ_* dependence, unlike \hat{q}_S .

Next, we investigate the variation of the p -value with the dimensionality of the look-elsewhere effect. To do this we extended the model to consider a signal at vector position \mathbf{x}_* . Each data point now corresponds to a vector \mathbf{x}^i , and we extend the signal and background in a symmetric fashion across each dimension, keeping the total prior volume fixed. Within the context of collider searches, the components of \mathbf{x}_* might correspond to a collection of invariant mass and jet properties. For astroparticle searches, the multiple dimensions might correspond to different directions in the sky. The bottom left plot of figure 4 shows the variation of the test statistics for dimensionality of 1, 2, and 3, for a constant prior volume of 100. It can be seen that, while the p -value of \hat{q}_L is dependent on the dimensionality, the p -value of \hat{q}_S is not. This justifies the naturally arising $(2\pi)^{M/2}$ prefactor in the posterior volume in equation (2.9). We also plot the $0d$ case, corresponding to only fitting for A with fixed x_* . Even though there is no look-elsewhere effect in this case, asymptotic agreement with equation (3.10) is still achieved. This shows our approach is still reliable in the $N \rightarrow 1$ limit, justifying its applicability for arbitrary N . As discussed in section 3.1, non-asymptotic agreement is not expected for a one-tailed test in the absence of the look-elsewhere effect, as the p -value tends to 0.5 as $\hat{q}_L \rightarrow 0$; on the other hand, a two-tailed test would give non-asymptotic agreement as shown in figure 2.

The above discussion concerns an un-binned model, parameterized by the signal fraction f . Often in particle physics, one performs a binned analysis with the number of events in each bin modelled as a Poisson distribution [35]. We find similar results when using this Poisson parameterization, as pictured in the bottom right of figure 4. The Poisson line agrees with the black line slightly better than the f line does, likely because the Laplace approximation is more accurate in the Poisson case.

When it comes to hypothesis testing, the relation between the true positive rate (TPR) and the false positive rate (FPR) determines the predictive power of a test statistic. In order to compare the relative power of the test statistics we consider an ROC plot for a variety of

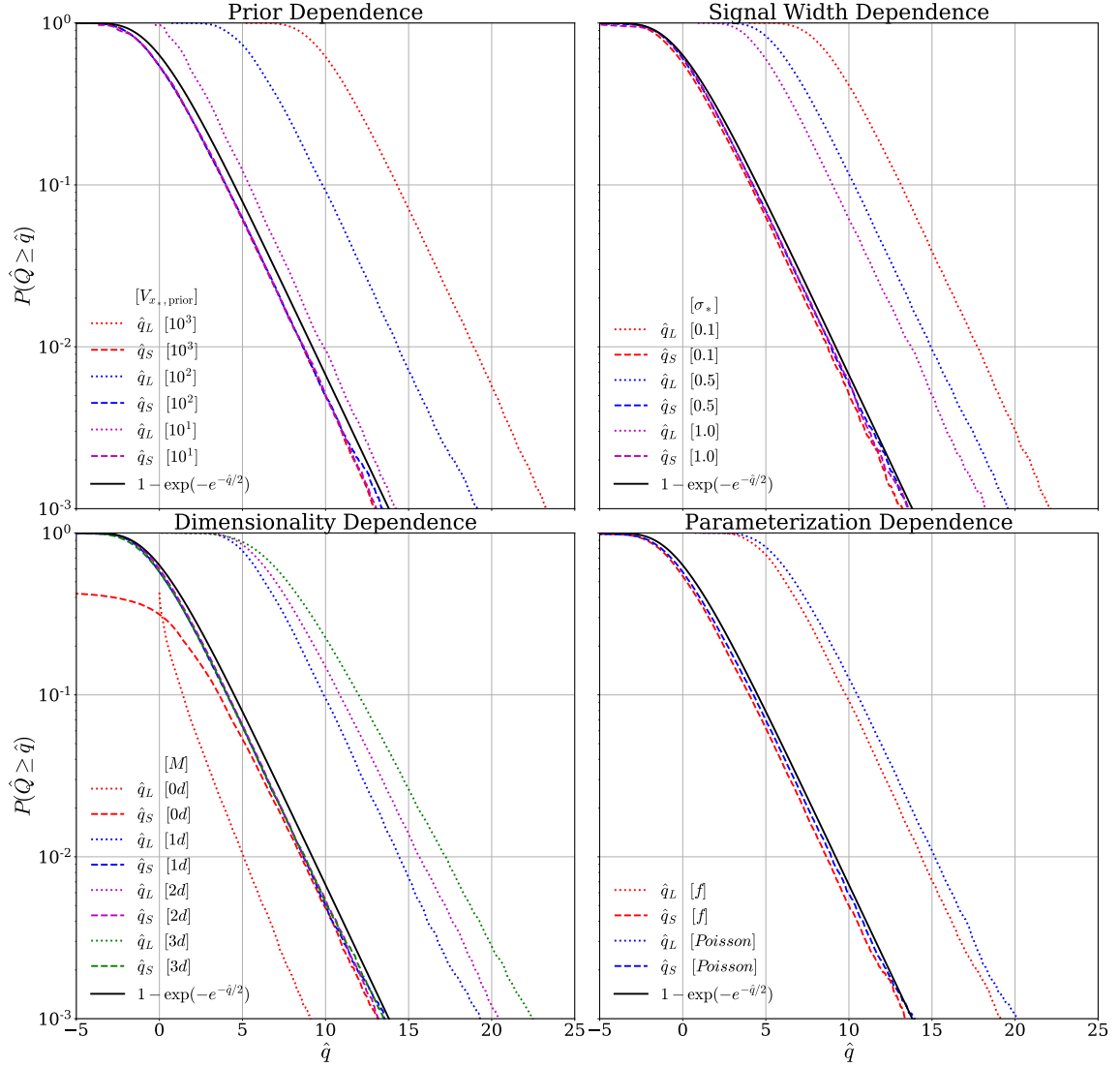


Figure 4. CCDFs of \hat{q}_L (dotted) and \hat{q}_S (dashed), computed using 10^5 simulations with no signal ($f = 0$). (Top Left) compares three prior volumes: 10^3 (red), 10^2 (blue), and 10^1 (magenta). (Top Right) compares different values of signal width σ_* : 0.1 (red), 0.5 (blue) and 1.0 (magenta). (Bottom Left) compares the dimensionality of \mathbf{x}_* : 0d (red), 1d (blue), 2d (magenta), and 3d (green). (Bottom Right) compares the un-binned f -parameterization (red) against a binned Poisson parameterization (blue). In all cases the p -value of \hat{q}_L has large variation, whereas \hat{q}_S does not. Furthermore, \hat{q}_S closely follows the predictions of equation (3.10) (black).

true f values, shown in figure 5. We also quote the (local) signal-to-noise ratio (SNR), which we define as the average $\sqrt{\hat{q}_L}$ across 10^4 realizations for the given f . It can be seen that \hat{q}_S and \hat{q}_L have approximately equivalent ROC lines, suggesting MAP and MPS have equal predictive power. This is expected as the relation between the test statistics is approximately monotonic, as seen in equation (3.3). This agrees with the findings of [36] which considered a different Bayesian-inspired test statistic, and showed it to have an approximately equivalent ROC curve to the p -value of the likelihood. Also, it can be seen that the predictive power increases with true f — as expected a larger true signal is more likely to be correctly detected.

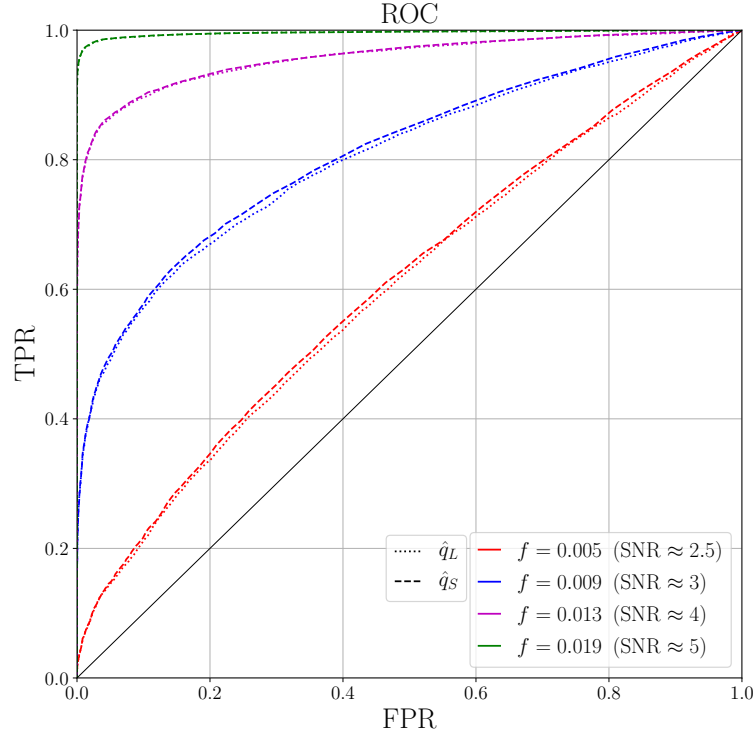


Figure 5. ROC curve: comparing the true positive rate (TPR), for a variety of f , with the false positive rate (FPR) for \hat{q}_L (dotted) and \hat{q}_S (dashed). The signal-to-noise ratio (SNR) corresponds to the average $\sqrt{\hat{q}_L}$ over all data realizations.

5 Example II: white noise

While we could continue the discussion in the context of resonance searches, we now consider a white noise time series example to illustrate the application of MPS to different models. This can be thought of as a toy model of a gravitational wave search. In this section we show how MPS handles additional model complexity as theorized in section 3.2. We consider a time series $y(x)$ comprising of measurements at N_d times, $\mathbf{x} = \{x^i\}_{i=1}^{N_d}$, with spacing $x^{i+1} - x^i = 1$. In the absence of a signal, each data point $y^i \equiv y(x^i)$ is assumed to be a standard normal random variable, i.e. we assume white noise. We consider a model with 2 degrees of freedom (dofs), with signal given by

$$p_s(x|A_1, A_2, x_*, \Delta, \sigma_*) = A_1 N(x|x_*, \sigma_*) + A_2 N(x|x_* + \Delta, \sigma_*) \quad (5.1)$$

where $A_{1,2} > 0$ are the amplitudes of each dof, x_* and $x_* + \Delta$ are the positions of the dofs, and σ_* is the common width.

As motivated in section 3.2, we reparameterize so that there's a single amplitude parameter, $z_1 = A$, and other parameters describing the properties of the single degree of freedom, $\mathbf{z}_{>1}$. We thus transform variables using $A_1 = A \cos \phi$ and $A_2 = A \sin \phi$, with $A > 0$ and $0 \leq \phi \leq \pi/2$ for a one-tailed test. By substituting the transformations into equation (5.1), the signal in the new parameterization is given by

$$p_s(x|A, \phi, x_*, \Delta, \sigma_*) = A [\cos \phi N(x|x_*, \sigma_*) + \sin \phi N(x|x_* + \Delta, \sigma_*)]. \quad (5.2)$$

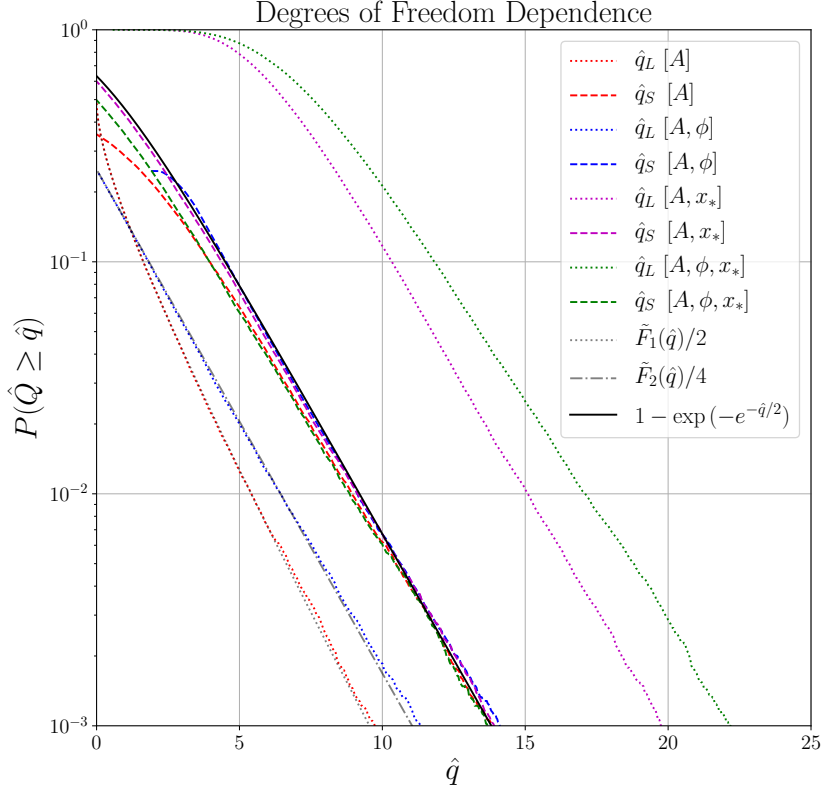


Figure 6. CCDFs of \hat{q}_L and \hat{q}_S averaged over 10^5 simulations with no signal ($A = 0$). The parameters in the square brackets are those being maximized, with other parameters being held fixed (as discussed in the text). The p -value of \hat{q}_L varies depending on the model complexity, whereas \hat{q}_S consistently follows the prediction of equation (3.10) (solid black).

The corresponding chi-squared difference between the data and the null hypothesis, equal to two times the log-likelihood-ratio, is given by

$$q_L(\mathbf{x}|A, \phi, x_*, \Delta, \sigma_*) = \sum_{i=1}^{N_d} [y^i - p_s(x^i|A, \phi, x_*, \Delta, \sigma_*)]^2 - [y^i]^2. \quad (5.3)$$

We consider a uniform prior on x_* with range $(0, 100)$, i.e. a prior volume of $V_{x_*, \text{prior}} = 100$, and $N_d = 100$. We do not fit for σ_* or Δ and fix them to $\sigma_* = 0.5$ and $\Delta = 10$. The application of MPS is identical to the previous section, so we will not repeat the methodology here.

Considering 10^5 data realizations with no signal, figure 6 shows how \hat{q}_L and \hat{q}_S are distributed for different levels of model complexity. First we maximize over A , while holding all other parameters fixed. In this case $\hat{q}_L \sim \tilde{F}_1(\hat{q}_L)/2$ (red dotted line) as expected for a one-tailed test with one degree of freedom. Additionally maximizing over ϕ allows for 2 dofs, and gives $\hat{q}_L \sim \tilde{F}_2(\hat{q}_L)/4$ (blue dotted line). This is expected because there are 4 permutations of each dof having positive or negative amplitude, and $A_{1,2} > 0$ considers 1 of these 4. For both of these cases, \hat{q}_S follows the same asymptotic distribution as predicted by equation (3.10). This verifies that the Bayesian picture of marginalizing over ϕ successfully reduces a model with 2 dofs to the same scale as 1 dof, in other words Wilks' Theorem has been replaced by marginalizing over ϕ . There is some discrepancy in the non-asymptotic regime for the maximization over A only (red dashed line), as discussed in section 3.1 for a one-tailed test.

We now introduce the look-elsewhere effect by allowing x_* to vary. First we maximize over A and x_* for fixed $\phi = 0$, as shown by the magenta lines. This is equivalent to a model with 1 dof because $\phi = 0$ corresponds to $A_2 = 0$. We see that the distribution of \hat{q}_L (magenta dotted line) is shifted to the right compared to the red and blue dotted lines due to the look-elsewhere effect. However, the distribution of \hat{q}_S (magenta dashed line) continues to follow the line predicted by equation (3.10). Finally, when maximizing over A , ϕ and x_* , i.e. a model with 2 dofs in the presence of the look-elsewhere effect, \hat{q}_L (green dotted line) is further right-shifted, whereas \hat{q}_S (green dashed line) again agrees with equation (3.10). The slight discrepancy in the A, ϕ, x_* maximization case is due to using too large a prior volume: there is a slight preference to having two well fitted peaks compared to one very well fitted peak, thus the distribution of ϕ is clustered towards $\phi = \pi/4$. Using a more appropriate prior for ϕ would improve agreement.

In summary, while the distribution of \hat{q}_L is highly dependent on the model complexity, via the extra degrees of freedom and look-elsewhere effect, \hat{q}_S has a universal distribution.

6 Example III: non-Gaussian models of cosmological inflation

There is much interest in detecting non-Gaussian models of inflation via the cosmological power spectrum [37–42]. A specific type of such a feature model adds the following oscillatory perturbation to the Λ CDM power spectrum,

$$P(k) = P_0(k)[1 + A \sin(2\omega k + \phi)], \quad (6.1)$$

where $P_0(k)$ is the featureless (Λ CDM) power spectrum and A , ω , and ϕ are the amplitude, frequency, and phase of the oscillatory perturbation. Such models are searched for using Planck 2013 data in [14] using the frequentist look-elsewhere analysis technique of [13]. In this section we seek to reproduce the conclusions of these papers using MPS.

Equation (6.1) can be written in the form $P(k) = P_0(k) + \Delta P(k)$ with

$$\begin{aligned} \Delta P(k; A, \omega, \phi) &= AP_0(k)[\cos \phi \sin(2\omega k) + \sin \phi \cos(2\omega k)] \\ &\equiv A \cos \phi P_s(k; \omega) + A \sin \phi P_c(k; \omega), \end{aligned} \quad (6.2)$$

where in the last line we explicitly separate terms with A and ϕ , as only ω couples to k . Assuming a linear relation, one can write $C_\ell = C_{\ell,0} + \Delta C_\ell$, with

$$\Delta C_\ell(A, \omega, \phi) = A \cos \phi C_{\ell,s}(\omega) + A \sin \phi C_{\ell,c}(\omega), \quad (6.3)$$

where $C_{\ell,s}$ and $C_{\ell,c}$ are the angular power spectra corresponding to P_s and P_c respectively. The Planck Likelihood [24] is given by

$$-2 \log L(\hat{C}_\ell | A, \omega, \phi) = [\hat{C}_{\ell_1} - C_{\ell_1}(A, \omega, \phi)] \Delta_{\ell_1 \ell_2} [\hat{C}_{\ell_2} - C_{\ell_2}(A, \omega, \phi)], \quad (6.4)$$

where \hat{C}_ℓ are the PCL estimates, and $\Delta_{\ell_1 \ell_2} = \langle \Delta \hat{C}_{\ell_1} \Delta \hat{C}_{\ell_2} \rangle$ is the PCL covariance matrix. In order to compute the likelihood for the null hypothesis, CosmoMC [43] was used to find the best fit values for the cosmological and nuisance parameters. When computing the likelihood for the signal hypothesis, the cosmological parameters were held fixed at these values; while they should really be re-fitted for the signal hypothesis, this is found to have little effect in [14]. The C_ℓ are evaluated using CAMB [44] with a sufficiently high accuracy setting to ensure resolution of the rapid oscillations. To speed up the evaluation of the likelihood over

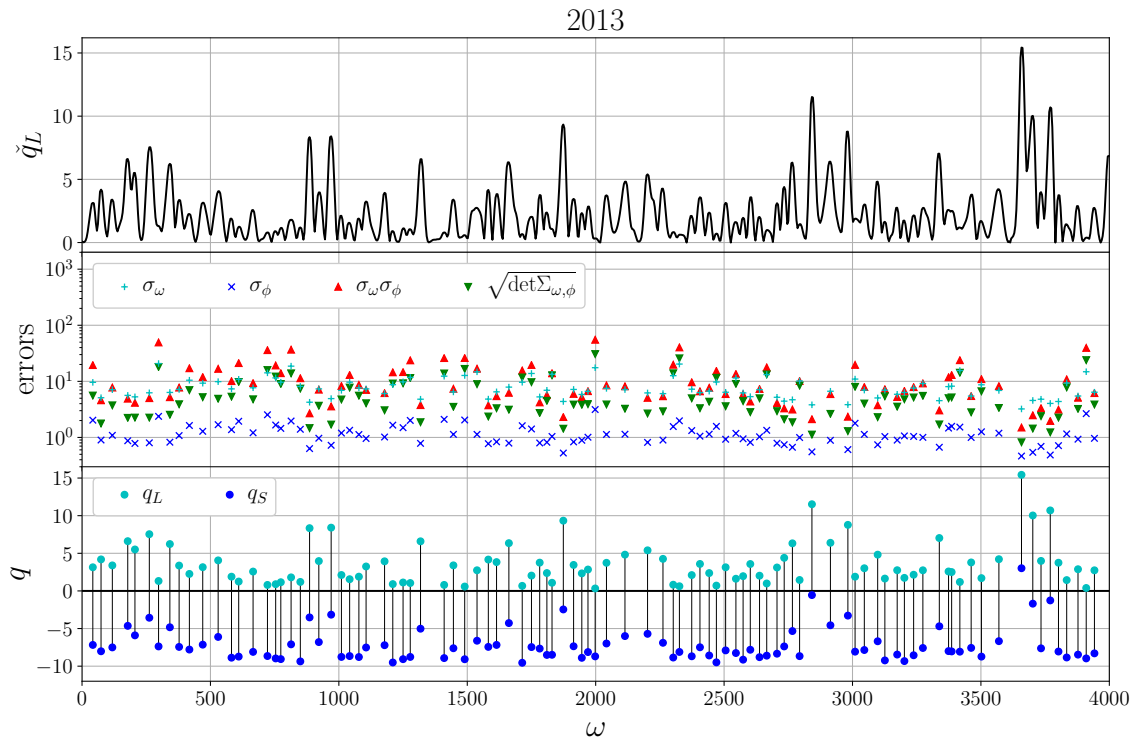


Figure 7. Planck results. Top: plot of \hat{q}_L , the projection of q_L onto the ω axis; this corresponds to q_L evaluated at the A and ϕ that maximize q_L at each ω . Middle: the errors obtained for the parameters, as well as a comparison with the determinant of the covariance matrix having removed the amplitude parameter, $\Sigma_{\omega,\phi}$. Bottom: a plot of q_L (blue) and q_S (cyan) for each peak, with the look-elsewhere correction depicted by the vertical black lines.

parameter space, $C_{\ell,s}(\omega)$ and $C_{\ell,c}(\omega)$ were computed over a discrete range of ω between 0 and 4000 with step-size $\Delta\omega = 5$, with intermediate values computed via spline interpolation. A flat prior was chosen for ω and ϕ . The rest of the analysis is analogous to the previous examples: we find all the local maxima of the posterior, compute the Hessian using finite differencing, compute the covariance matrix, and use this to find \hat{q}_S . Unlike the previous examples, we note that ω and ϕ are correlated, as illustrated in the middle plot of figure 7, so it is important to use the determinant of the full covariance matrix and not just its diagonal components. It is also interesting to note that higher peaks have smaller errors.

The results obtained using the CAMspec component of the 2013 Planck likelihood² are pictured in figure 7. The maximum occurs at $\omega \approx 3660$ with $\hat{q}_L = 15.4$, giving a naive significance of $\sqrt{\hat{q}_L} \approx 4$ sigma. However, we find that $\hat{q}_S = 3.0$, giving a global p -value of $1 - \exp(1 - e^{-3/2}) = 0.20$ using equation (3.10), and significance of $S = 1.3$ sigma. Thus the signal is in fact far less significant in light of the look-elsewhere effect. The prescription of [14] gives a p -value of 0.13, which is in reasonable agreement. Note that our likelihood profile does not match [14] exactly due to our approximate approach, hence the p -value quoted here is the value one would obtain by applying the prescription of [14] to our likelihood profile. We applied the same analysis to the 2015 `plik_lite` likelihood [45] and found a p -value of approximately 1, suggesting no evidence for such models of non-Gaussianity.

²One should sum the different components of the likelihood, but this is unnecessary for our proof of concept.

7 Conclusions

This work has employed Bayesian and frequentist thinking to provide a fast method to account for the look-elsewhere effect when scanning over a large parameter space. We started by considering the Bayesian approach, and explained how maximizing the posterior mass, as in MPM, is a more appropriate choice than maximizing the posterior density, as in MAP. Bayesian methodology naturally considers model complexity and the look-elsewhere effect by marginalization, which penalizes the likelihood by the prior-to-posterior volume ratio. (Under the Laplace approximation, the posterior volume is simply related to the determinant of the covariance matrix.) We then considered the frequentist approach by writing the global p -value as the local p -value multiplied by the trials factor. By drawing an analogy between the two approaches we identified the trials factor as the prior-to-posterior volume ratio of the parameters being scanned over, in turn generalizing the Bonferroni correction to continuous problems. We introduced q_S and in turn MPS, a hybrid of MPM and MAP, which considers the mode with maximum q_S . Finally, we generalized the Šidák correction to continuous problems, providing a universal way to assign the global p -value in both the asymptotic and non-asymptotic regimes.

We illustrated the effectiveness of MPS by considering several examples from (astro)particle physics and cosmology, showing it to have equal predictive power to MAP while naturally accounting for the look-elsewhere effect. MPS effectively shifts the hypothesis testing threshold of the maximum likelihood ratio to a generic scale: while the maximum likelihood ratio, or equivalently the best fit chi-squared $\chi^2 = \hat{q}_L$, depends on the model complexity and extent of the look-elsewhere effect, \hat{q}_S does not. In other words, instead of considering fixed \hat{q}_L thresholds, one should consider fixed \hat{q}_S thresholds.

Unlike current methods that rely on performing numerous simulations, MPS accounts for the look-elsewhere effect by using information from the data alone, as one need only compute the likelihood and the posterior volume to evaluate q_S . This provides a more efficient way to quantify statistical significance as it does not require expensive simulations. In a typical situation one would focus on the most promising anomalies only, with \hat{q}_S providing a scale that gives good guidance on what false positive rate one should expect. Subsequently, one would obtain additional information to verify the veracity of an anomaly when possible.

For our proof of concept it was sufficient to only consider simple physical examples in this paper, but there are many real-world applications where our methods can be employed. Examples include searches for new particles in astroparticle and particle data, searches for gravitational wave signals in LIGO/VIRGO data, searches for exoplanets in transit and radial velocity data, as well as many more. In some of these cases the look-elsewhere penalty can be considerably large, reaching beyond 6 sigma. The problem is very general, as almost every search for unknown objects, events, new physics, or other phenomena whose existence is unknown, has to deal with the look-elsewhere effect. We note that while this work only explored the look-elsewhere effect from scanning over parameter space, we expect that similar methods can be applied to other manifestations of the look-elsewhere effect, for example scanning over models when fitting gravitational wave templates.

The goal of a data analyst searching for anomalies is to report the most promising anomalies in terms of having a small p -value, or a high Bayes factor. By clarifying the origins of the look-elsewhere effect and model complexity penalty for continuous parameters we hope to open the way to refinements in anomaly searches that can improve the overall success rate of a detection. This should be a common goal of any experimental analysis regardless of which school of statistics one belongs to.

Acknowledgments

We thank Benjamin Nachman for insightful comments on the manuscript, and Benjamin Wallisch for valuable discussion regarding example III. This research made use of the Cori supercomputer at the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This material is based upon work supported by the National Science Foundation under Grant Numbers 1814370 and NSF 1839217, and by NASA under Grant Number 80NSSC18K1274.

A Derivation of the CCDF of \hat{q}_S

The asymptotic (large q_L) CCDF of the global maximum of q_L is for a one-tail test is given in equation (3.2) as

$$P_{Q_L}(Q_L \geq q_L) = N \frac{1}{2} \tilde{F}_1(q_L) \quad (\text{A.1})$$

$$= N \frac{1}{\sqrt{2\pi q_L}} e^{-q_L/2} + N q_L^{-1/2} e^{-q_L/2} \mathcal{O}(q_L^{-1}), \quad (\text{A.2})$$

where here we include the leading order correction, and drop hats and take $t = 1$ for convenience. Consider the transformation of variables to q_S , defined by

$$q_S \equiv g(q_L) \equiv q_L - 2 \ln N + \ln 2\pi q_L. \quad (\text{A.3})$$

It can be shown that the inverse of g is given by

$$q_L = g^{-1}(q_S) = W_0\left(\frac{N^2 e^{q_S}}{2\pi}\right) \quad (\text{A.4})$$

$$= q_S + \ln \frac{N^2}{2\pi} - \ln\left(q_S + \ln \frac{N^2}{2\pi}\right) + \mathcal{O}\left(\frac{L_2}{L_1}\right), \quad (\text{A.5})$$

where $W_0(z)$ is the principal branch of the Lambert W function. The asymptotic expansion has been performed in the final line, with the shorthand $L_i \equiv \ln^i \frac{N^2 e^{q_S}}{2\pi}$. Assuming N is constant to study the limiting behaviour, the CCDF of q_S is thus

$$P_{Q_S}(Q_S \geq q_S) = P_{Q_L}[Q_L \geq g^{-1}(q_S)] \quad (\text{A.6})$$

$$= e^{-q_S/2} e^{-\mathcal{O}(L_2/L_1)} \left(1 - \frac{\ln\left(q_S + \ln \frac{N^2}{2\pi}\right) + \mathcal{O}\left(\frac{L_2}{L_1}\right)}{q_S + \ln \frac{N^2}{2\pi}}\right)^{-1/2} + \mathcal{O}\left(\frac{e^{-q_S/2}}{q_S + \ln \frac{N^2}{2\pi}}\right) \quad (\text{A.7})$$

$$\rightarrow e^{-q_S/2}, \quad (\text{A.8})$$

where the limit corresponds to either $N \rightarrow \infty$ or $q_S \rightarrow \infty$. This means the result still applies asymptotically in the absence of the look-elsewhere effect ($N = 1$).

References

- [1] R.G. Miller, *Simultaneous Statistical Inference*, Springer, New York U.S.A. (1981).
- [2] J.P. Shaffer, *Multiple hypothesis testing*, *Ann. Rev. Psychol.* **46** (1995) 561.

- [3] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1 [[arXiv:1207.7214](#)] [[INSPIRE](#)].
- [4] CMS collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30 [[arXiv:1207.7235](#)] [[INSPIRE](#)].
- [5] B. Anderson, S. Zimmer, J. Conrad, M. Gustafsson, M. Sánchez-Conde and R. Caputo, *Search for Gamma-Ray Lines towards Galaxy Clusters with the Fermi-LAT*, *JCAP* **02** (2016) 026 [[arXiv:1511.00014](#)] [[INSPIRE](#)].
- [6] A. Reinert and M.W. Winkler, *A Precision Search for WIMPs with Charged Cosmic Rays*, *JCAP* **01** (2018) 055 [[arXiv:1712.00002](#)] [[INSPIRE](#)].
- [7] N. Sekiya, N.Y. Yamasaki and K. Mitsuda, *A search for a keV signature of radiatively decaying dark matter with Suzaku XIS observations of the X-ray diffuse background*, *Publ. Astron. Soc. Jpn.* **68** (2015) S31 [<https://academic.oup.com/pasj/article-pdf/68/SP1/S31/7971976/psv081.pdf>].
- [8] ICECUBE collaboration, *Observation of High-Energy Astrophysical Neutrinos in Three Years of IceCube Data*, *Phys. Rev. Lett.* **113** (2014) 101101 [[arXiv:1405.5303](#)] [[INSPIRE](#)].
- [9] K. Emig, C. Lunardini and R. Windhorst, *Do high energy astrophysical neutrinos trace star formation?*, *JCAP* **12** (2015) 029 [[arXiv:1507.05711](#)] [[INSPIRE](#)].
- [10] K. Cannon, C. Hanna and J. Peoples, *Likelihood-ratio ranking statistic for compact binary coalescence candidates with rate estimation*, [arXiv:1504.04632](#).
- [11] LIGO SCIENTIFIC and VIRGO collaborations, *GW150914: First results from the search for binary black hole coalescence with Advanced LIGO*, *Phys. Rev. D* **93** (2016) 122003 [[arXiv:1602.03839](#)] [[INSPIRE](#)].
- [12] C. Messick et al., *Analysis Framework for the Prompt Discovery of Compact Binary Mergers in Gravitational-wave Data*, *Phys. Rev. D* **95** (2017) 042001 [[arXiv:1604.04324](#)] [[INSPIRE](#)].
- [13] J.R. Fergusson, H.F. Gruetjen, E.P.S. Shellard and M. Liguori, *Combining power spectrum and bispectrum measurements to detect oscillatory features*, *Phys. Rev. D* **91** (2015) 023502 [[arXiv:1410.5114](#)] [[INSPIRE](#)].
- [14] J.R. Fergusson, H.F. Gruetjen, E.P.S. Shellard and B. Wallisch, *Polyspectra searches for sharp oscillatory features in cosmic microwave sky data*, *Phys. Rev. D* **91** (2015) 123506 [[arXiv:1412.6152](#)] [[INSPIRE](#)].
- [15] P. Hunt and S. Sarkar, *Search for features in the spectrum of primordial perturbations using Planck and other datasets*, *JCAP* **12** (2015) 052 [[arXiv:1510.03338](#)] [[INSPIRE](#)].
- [16] J. Robnik and U. Seljak, *Kepler data analysis: non-gaussian noise and fourier gaussian process analysis of star variability*, [arXiv:1910.01167](#).
- [17] P.I.W. de Bakker et al., *Efficiency and power in genetic association studies*, *Nat. Genet.* **37** (2005) 1217.
- [18] J.D. Storey and R. Tibshirani, *Statistical significance for genomewide studies*, *Proc. Natl. Acad. Sci. U.S.A.* **100** (2003) 9440.
- [19] M. Proschan and M. Waclawiw, *Practical guidelines for multiplicity adjustment in clinical trials*, *Control. Clin. Trials* **21** (2001) 527.
- [20] B. McKay, D. Bar-Natan, M. Bar-Hillel and G. Kalai, *Solving the bible code puzzle*, *Statist. Sci.* **14** (1999) 150.
- [21] C.E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilità*, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, Florence Italy (1936).
- [22] Z. Šidák, *Rectangular confidence regions for the means of multivariate normal distributions*, *J. Am. Statist. Assoc.* **62** (1967) 626.

- [23] E. Gross and O. Vitells, *Trial factors for the look elsewhere effect in high energy physics*, *Eur. Phys. J. C* **70** (2010) 525 [[arXiv:1005.1891](#)] [[INSPIRE](#)].
- [24] PLANCK collaboration, *Planck 2013 results. XV. CMB power spectra and likelihood*, *Astron. Astrophys.* **571** (2014) A15 [[arXiv:1303.5075](#)] [[INSPIRE](#)].
- [25] M.-H. Chen, Q.-M. Shao and J.G. Ibrahim, *Estimating Ratios of Normalizing Constants*, in *Monte Carlo Methods in Bayesian Computation*, Springer, New York U.S.A. (2000), pg. 124.
- [26] H. Jia and U. Seljak, *Normalizing constant estimation with gaussianized bridge sampling*, [arXiv:1912.06073](#) [[INSPIRE](#)].
- [27] U. Seljak and B. Yu, *Posterior inference unchained with EL_{2O}*, [arXiv:1901.04454](#) [[INSPIRE](#)].
- [28] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge U.K. (2003).
- [29] D.V. Lindley, *A statistical paradox*, *Biometrika* **44** (1957) 187.
- [30] R.D. Cousins, *The Jeffreys-Lindley paradox and discovery criteria in high energy physics*, *Synthese* **194** (2017) 395 [[arXiv:1310.3791](#)] [[INSPIRE](#)].
- [31] S. Nesseris and J. García-Bellido, *Is the Jeffreys' scale a reliable tool for Bayesian model comparison in cosmology?*, *JCAP* **08** (2013) 036 [[arXiv:1210.7652](#)] [[INSPIRE](#)].
- [32] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, *Ann. Math. Statist.* **9** (1938) 60.
- [33] R.B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, *Biometrika* **64** (1977) 247.
- [34] R.B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, *Biometrika* **74** (1987) 33.
- [35] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J. C* **71** (2011) 1554 [Erratum *ibid.* **73** (2013) 2501] [[arXiv:1007.1727](#)] [[INSPIRE](#)].
- [36] A. Fowlie, *Bayesian and frequentist approaches to resonance searches*, *2019 JINST* **14** P10031 [[arXiv:1902.03243](#)].
- [37] J. Maldacena, *Non-gaussian features of primordial fluctuations in single field inflationary models*, *JHEP* **05** (2003) 013.
- [38] P. Creminelli, A. Nicolis, L. Senatore, M. Tegmark and M. Zaldarriaga, *Limits on non-Gaussianities from wmap data*, *JCAP* **05** (2006) 004 [[astro-ph/0509029](#)] [[INSPIRE](#)].
- [39] E. Komatsu et al., *Five-year wilkinson microwave anisotropy probe observations: Cosmological interpretation*, *Astrophys. J. Suppl. Ser.* **180** (2009) 330.
- [40] U. Seljak, *Extracting primordial non-Gaussianity without cosmic variance*, *Phys. Rev. Lett.* **102** (2009) 021302 [[arXiv:0807.1770](#)] [[INSPIRE](#)].
- [41] PLANCK collaboration, *Planck 2013 Results. XXIV. Constraints on primordial non-Gaussianity*, *Astron. Astrophys.* **571** (2014) A24 [[arXiv:1303.5084](#)] [[INSPIRE](#)].
- [42] PLANCK collaboration, *Planck 2018 results. IX. Constraints on primordial non-Gaussianity*, [arXiv:1905.05697](#) [[INSPIRE](#)].
- [43] A. Lewis and S. Bridle, *Cosmological parameters from CMB and other data: A Monte Carlo approach*, *Phys. Rev. D* **66** (2002) 103511 [[astro-ph/0205436](#)] [[INSPIRE](#)].
- [44] A. Lewis, A. Challinor and A. Lasenby, *Efficient computation of CMB anisotropies in closed FRW models*, *Astrophys. J.* **538** (2000) 473 [[astro-ph/9911177](#)] [[INSPIRE](#)].
- [45] PLANCK collaboration, *Planck 2015 results. XI. CMB power spectra, likelihoods, and robustness of parameters*, *Astron. Astrophys.* **594** (2016) A11 [[arXiv:1507.02704](#)] [[INSPIRE](#)].