

Bayesian evidence and model selection



Kevin H. Knuth^{a,b,*}, Michael Habeck^{c,d}, Nabin K. Malakar^e, Asim M. Mubeen^{a,f},
Ben Placek^a

^a Dept. of Physics, Univ. at Albany (SUNY), Albany NY 12222, USA

^b Dept. of Informatics, Univ. at Albany (SUNY), Albany NY 12222, USA

^c Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

^d Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen, 37077 Göttingen, Germany

^e Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA 91109, USA

^f Geriatrics Division, Nathan Kline Institute, Orangeburg NY 10962, USA

ARTICLE INFO

Article history:

Available online 29 June 2015

Keywords:

Bayesian signal processing
Bayesian evidence
Model testing
Nested sampling
Odds ratio

ABSTRACT

In this paper we review the concepts of Bayesian evidence and Bayes factors, also known as log odds ratios, and their application to model selection. The theory is presented along with a discussion of analytic, approximate and numerical techniques. Specific attention is paid to the Laplace approximation, variational Bayes, importance sampling, thermodynamic integration, and nested sampling and its recent variants. Analogies to statistical physics, from which many of these techniques originate, are discussed in order to provide readers with deeper insights that may lead to new techniques. The utility of Bayesian model testing in the domain sciences is demonstrated by presenting four specific practical examples considered within the context of signal processing in the areas of signal detection, sensor characterization, scientific model selection and molecular force characterization.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The application of model-based reasoning techniques employing Bayesian probability theory has recently found wide use in signal processing, and in the physical sciences in general [1–6]. In such an approach, it is critical to be able to statistically compare the probability of one model to another. This is performed by computing the Bayesian evidence of the two models and comparing them by forming a ratio, which is often referred to as a Bayes factor or the odds ratio.

In this paper, we present an overview of the theory behind Bayesian evidence, discuss various methods of computation, and demonstrate the application in four practical examples of current interest more closely related to signal processing. We do not aim to cover all of the techniques and applications, as there exists a great number of excellent treatments spanning several decades [7–9,1,10–13,2,14,5,15–17,4] as well as a wide variety of applications spread across a great number of fields, such as acoustics [18–20], astronomy, astrophysics and cosmology [21–31], chemistry [32], computer science and machine learning [33,34], neu-

ral networks [35–37], neuroscience [38–41], nuclear and particle physics [42–44], signal processing [45–51] systems engineering [52–54], and statistics in general [55,56].

2. Probability

Logical statements can imply other logical statements. Probability theory [57,58,1,16,2,17,59,14,5,6] allows one to generalize the concept of implication by providing a measure of the degree of implication among logical statements [60,61]. More specifically, probability is a scalar measure that quantifies, within a topic of discourse, the degree to which one logical statement, representing a state of knowledge, implies another [62,63].¹ As a scalar measure, probability enables one to rank logical statements with respect to a given context or premise.

The utility of probability theory becomes apparent when one considers the degree to which a statement considering a set of several hypotheses or models, M , implies a joint statement proposing a particular model m in conjunction with additional information or data, d , which we write as $P(m, d|M)$. The product rule, which can

* Corresponding author at: Dept. of Physics, Univ. at Albany (SUNY), Albany NY 12222, USA.

E-mail address: kknuth@albany.edu (K.H. Knuth).

¹ This is a relatively new interpretation of probability that has significant advantages over older concepts such as the *frequency of occurrences of events*, the *degree of truth* or the *degree of belief*.

be derived as a consequence of basic symmetries of Boolean logic [60–63], enable one to express this probability in two ways

$$P(m, d|M) = P(m|M)P(d|m, M) \quad (1)$$

$$= P(d|M)P(m|d, M). \quad (2)$$

These two expressions can be equated

$$P(m|M)P(d|m, M) = P(d|M)P(m|d, M) \quad (3)$$

and rearranged resulting in the familiar Bayes' theorem

$$P(m|d, M) = P(m|M) \frac{P(d|m, M)}{P(d|M)}, \quad (4)$$

where the posterior probability $P(m|d, M)$ can be expressed in terms of the product of the prior probability $P(m|M)$ with a data-dependent term consisting of the ratio of the likelihood $P(d|m, M)$ to the evidence $P(d|M)$. It is in this sense that one can think of Bayes' theorem as a learning rule where one's prior state of knowledge about the problem, represented by the prior probability, is updated by a data-dependent term resulting in a posterior probability that depends both on the prior state of knowledge as well as the data.

Both the prior probability and the likelihood must be assigned based on any and all additional information that one may possess about the problem. This is not a deficit or drawback of probability theory. Instead it is a strength since symmetries only serve to constrain manipulation of probabilities to the sum and product rules. This leaves free the probability assignments resulting in a theory of inductive logic that can be applied to any particular inference problem. The dependence of these probabilities on problem-specific prior information is often indicated by including the symbol I to the right of the solidus.² For example, this is done by writing the prior probability $P(m|M)$ as $P(m|M, I)$.

While the posterior probability over the space of models M fully quantifies all that is known about the problem, it is often common practice to summarize what is known by focusing on a particular model m that maximizes the posterior probability, such that this model is most implied by the data given the prior information. Such a model is referred to as the most probable model or mode (within the context defined by the space of models M), or the maximum a posteriori (MAP) estimate. Often the space of models M to be considered is a parameterized space where each model m is represented by a set of particular parameter values that act as coordinates in the space. In this case, one can consider summarizing the posterior using the model given by the mean parameter values found using the posterior. Either way, when the models in the space M are parameterized, selecting a particular model given the data and prior information amounts to a parameter estimation problem.

The evidence, which in parameter estimation problems acts mainly as a normalization factor, can be obtained by summing or integrating (marginalizing) over all possible models m in the set of models M

$$P(d|M, I) = \int dm P(m, d|M, I) \quad (5)$$

$$= \int dm P(m|M, I)P(d|m, M, I), \quad (6)$$

which is the reason that the evidence is often referred to as the marginal likelihood.

We can refer to a set of models, M , as a particular theory. Given two competing theories M_1 or M_2 one can compare the posterior probability $P(M_1|d, I)$ to the posterior probability $P(M_2|d, I)$, where, among additional prior information, I represents the fact that theories M_1 and M_2 are among those to be considered. In general, both theories will result in non-zero probabilities. However, the more probable theory can be determined by considering the ratio of their posterior probabilities. We can examine this by considering the ratio of joint probabilities of the sets of models M_1 and M_2 and the data d and then using the product rule to write the joint probability in two ways

$$\frac{P(M_1, d|I)}{P(M_2, d|I)} = \frac{P(M_1, d|I)}{P(M_2, d|I)} \quad (7)$$

$$\frac{P(d|I)P(M_1|d, I)}{P(d|I)P(M_2|d, I)} = \frac{P(M_1|I)P(d|M_1, I)}{P(M_2|I)P(d|M_2, I)} \quad (8)$$

$$\frac{P(M_1|d, I)}{P(M_2|d, I)} = \frac{P(M_1|I)}{P(M_2|I)} \frac{P(d|M_1, I)}{P(d|M_2, I)} \quad (9)$$

so that the ratio of the posterior probabilities of the two theories is proportional to the ratio of their respective evidences. The proportionality becomes an equality in the case where the prior probabilities of the two theories are equal. This leads to the concept of the Bayes factor or odds ratio where we define

$$\text{OR} = \frac{P(d|M_1, I)}{P(d|M_2, I)} \quad (10)$$

or, equivalently, the log odds ratio

$$\log \text{OR} = \log P(d|M_1, I) - \log P(d|M_2, I). \quad (11)$$

With this definition, we can write the ratio of posterior probabilities for the two different theories M_1 and M_2 in terms of the odds ratio

$$\frac{P(M_1|d, I)}{P(M_2|d, I)} = \frac{P(M_1|I)}{P(M_2|I)} \times \text{OR}, \quad (12)$$

where the two are equal when the ratio of the prior probabilities of the two theories are equal.

In the case of parameter estimation problems, the Bayesian evidence plays a relatively minor role as a normalization factor. However, in problems where two theories are being tested against one another, which is often called a model selection problem,³ the ratio of evidences is the relevant quantity to consider. In some special cases, the integrals can be solved analytically as described in [8,10] and demonstrated below in Section 5.1.

3. Evidence, model order, and priors

It is instructive to consider how the evidence (6) varies as a function of the considered model order as well as the prior information one may possess about the model. We begin by considering a model consisting of a single parameter x , for which we have assigned a uniform prior probability over an interval $[x_{\min}, x_{\max}]$ of width $\Delta x = x_{\max} - x_{\min}$. We can define the effective width $\delta x \leq \Delta x$ of the likelihood over the prior range as

$$\delta x \doteq \frac{1}{L_{\max}} \int_{x_{\min}}^{x_{\max}} dx P(d|x, M, I), \quad (13)$$

² This notation goes back to Jaynes [1] and has been adopted in several prominent textbooks in the physical sciences [64,17,2,6].

³ The terminology may be confusing since the term 'model selection' seems to refer to the process of selecting a particular model; whereas, it refers to selecting one set of models, or theory, over another.

where L_{\max} is the value of the likelihood $P(d|x, M, I)$ attained at the maximum likelihood estimate $x = \hat{x}$. The evidence of the model amounts to

$$Z \equiv P(d|M, I) = \frac{1}{\Delta x} \int_{x_{\min}}^{x_{\max}} dx P(d|x, M, I), \quad (14)$$

which using the definition in (13) can be conveniently expressed in terms of the prior width Δx and effective likelihood width δx by [64, pp. 63–65]

$$Z = L_{\max} \frac{\delta x}{\Delta x}. \quad (15)$$

Thus we can write the evidence as a product of the maximum of the likelihood (the best achievable goodness-of-fit) and an Occam factor W :

$$Z = L_{\max} W \quad (16)$$

where $0 \leq W \leq 1$ is formally defined as

$$W = \frac{Z}{L_{\max}} = \int dx P(x|M, I) \frac{P(d|x, M, I)}{L_{\max}}. \quad (17)$$

For models with a single adjustable parameter the Occam factor is the ratio of the width of the likelihood over the prior range to the width of the prior: $W = \delta x / \Delta x$. For multiple model parameters this generalizes to the ratio of the volume occupied by those models that are compatible with both data and prior over the prior accessible volume.

By making the prior broader we pay in evidence. It is in this sense that Bayesian probability theory embodies Occam's razor: "Entities are not to be multiplied without necessity." If we increase the flexibility of our model by the introduction of more model parameters, we reduce the Occam factor. Let's for simplicity assume that every additional parameter is also uniform over an interval of length Δx and that there are K such parameters x_k . Then beyond a certain model order K , we will achieve a perfect fit of the data upon which we cannot improve the likelihood any further. Because the Occam factor scales as $(\delta x / \Delta x)^K$, it will disfavor a further increase in model order.

Consider a Gaussian likelihood function, which is normalized so that it integrates to unity. If the data $d = \{d_1, \dots, d_n\}$ are modeled as independent observations, the likelihood, assuming a standard deviation σ , is

$$P(d|x, M, I) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}[(x - \bar{d})^2 + v]\right\} \quad (18)$$

where $\bar{d} = \frac{1}{n} \sum_i d_i$ is the sample average and $v = \frac{1}{n} \sum_i (d_i - \bar{d})^2$ the sample variance. Maximum likelihood is obtained at $\hat{x} = \bar{d}$ achieving a likelihood of $L_{\max} = (e^{-v/\sigma^2} / 2\pi\sigma^2)^{n/2}$. The evidence is

$$P(d|M, I) = L_{\max} \sqrt{\frac{2\pi}{n}} \frac{\sigma}{\Delta x} \frac{\operatorname{erf}\left(\sqrt{\frac{n}{2}} \frac{x_{\max} - \bar{d}}{\sigma}\right) + \operatorname{erf}\left(\sqrt{\frac{n}{2}} \frac{\bar{d} - x_{\min}}{\sigma}\right)}{2}. \quad (19)$$

For $\bar{d} \in [x_{\min}, x_{\max}]$ and σ small or n large, we can ignore the last factor involving the error function. The Occam factor is essentially $\sqrt{2\pi}\sigma / \Delta x \sqrt{n}$. If \bar{d} falls outside the support of the prior ($\bar{d} < x_{\min}$ or $\bar{d} > x_{\max}$), the evidence decreases rapidly reflecting the discrepancy between our prior assumptions and the actual observations.

Let us compare a model M_0 that has no adjustable parameter and a model M_1 with a single adjustable parameter x by computing the odds ratio:

$$\text{OR} = \frac{P(d|M_0, I)}{P(d|M_1, I)} \approx \frac{P(D|M_0, I)}{P(D|\hat{x}, M_1, I)} \frac{\Delta x}{\delta x} \quad (20)$$

The odds ratio is comprised of two factors: the ratio of the likelihoods

$$\frac{P(D|M_0, I)}{P(D|\hat{x}, M_1, I)}$$

and the Occam factor $\Delta x / \delta x$. The likelihood ratio is a classical statistic in frequentist model selection. If we only consider the likelihood ratio in model comparison problems, we fail to acknowledge the importance of Occam factors.

4. Numerical techniques

In general, the evidence, which is found by integrating the prior times the likelihood (6) over the entire parameter space, cannot be solved analytically.⁴ This requires that we use numerical techniques to estimate the evidence. Straightforward estimation of the evidence integral directly from posterior sampling, such as in [65], proves to be quite challenging in general, especially in the case of multimodal distributions arising from mixture models or high-dimensional spaces. While a number of sophisticated problem-specific techniques have been developed to handle such difficulties [66–68], there is a need for more general widely-applicable techniques that require little to no fine tuning.

Other methods, such as *Reversible Jump Markov Chain Monte Carlo* (RJMCMC) treat the model order as a model parameter [69–71]. However, such techniques typically encounter serious difficulties with inefficient model-switching moves. The difficulties these more direct techniques experience are especially problematic in high-dimensional spaces and in problems where the likelihood calculations are expensive, such as in the case of large data sets or complex forward models.

This has resulted in the development of a rather sophisticated array of computational techniques. Here we briefly review some of the more popular methods, pointing the interested readers to additional excellent resources and reviews, such as [9] and [72], and conclude with a focus on the more recent methods of nested sampling and its cousin MultiNest, which are used in three of the examples provided in the following section.

4.1. Laplace approximation

The *Laplace Approximation*, also known as the *Saddle-Point Approximation* [73], is a simple and useful method for approximating a unimodal probability density function with a Gaussian [16,74,4,6]. As such, the Laplace approximation forms the basis of more advanced techniques, such as Gull and MacKay's *Evidence Framework* [75,35].

Consider a function $p(x)$, which has a peak at $x = x_0$. One can write the Taylor series expansion of the logarithm of the probability density $\ln p(x)$ about $x = x_0$ to second order as

$$\begin{aligned} \ln p(x) \simeq \ln p(x_0) + \frac{d}{dx} \ln p(x) \Big|_{x=x_0} (x - x_0) \\ + \frac{1}{2} \frac{d^2}{dx^2} \ln p(x) \Big|_{x=x_0} (x - x_0)^2 + \dots, \end{aligned} \quad (21)$$

which can be simplified to

$$\ln p(x) \simeq \ln p(x_0) + \frac{1}{2} \frac{d^2}{dx^2} \ln p(x) \Big|_{x=x_0} (x - x_0)^2 + \dots, \quad (22)$$

⁴ A rare exception is given by the first example presented in Section 5.1 where an analytical solution is obtained.

since the first derivative of $\ln p(x)$ evaluated at the peak is zero. By defining σ^2 to be minus the inverse of the local curvature at the peak

$$\sigma^2 = \left(-\frac{1}{2} \frac{d^2}{dx^2} \ln p(x) \Big|_{x=x_0} \right)^{-1}, \quad (23)$$

we can rewrite (22) as

$$\ln p(x) \simeq \ln p(x_0) - \frac{1}{2\sigma^2} (x - x_0)^2 + \dots \quad (24)$$

Taking the exponential of both sides results in an un-normalized approximation for $p(x)$

$$p(x) \simeq p(x_0) \exp \left[-\frac{1}{2\sigma^2} (x - x_0)^2 \right], \quad (25)$$

which would have as its normalization factor

$$Z = p(x_0) \sqrt{2\pi\sigma^2}. \quad (26)$$

If the function $p(x)$ is taken to be the product of the prior probability and the likelihood, then, the normalization factor (26) is an approximation of the evidence.

In N dimensions, we expand the function $\ln p(\mathbf{x})$ as

$$\ln p(\mathbf{x}) \simeq \ln p(\mathbf{x}_0) - \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \mathbf{A} (\mathbf{x} - \mathbf{x}_0) + \dots, \quad (27)$$

where \mathbf{A} is an $N \times N$ matrix, known as the Hessian, with matrix elements given by

$$A_{ij} = -\frac{d^2}{dx_i dx_j} \ln p(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0}. \quad (28)$$

The approximation of $p(\mathbf{x})$ is then given by

$$p(\mathbf{x}) \simeq \frac{1}{Z} \exp \left[-\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} \right] \quad (29)$$

where the normalization factor is

$$Z = p(\mathbf{x}_0) \sqrt{\frac{(2\pi)^N}{\det \mathbf{A}}}. \quad (30)$$

Again, if the function $p(\mathbf{x})$ is defined by the product of the prior and the likelihood, then Z is the approximation to the evidence. This method requires that the peak of the distribution be identified and the Hessian estimated either analytically or numerically.

The Laplace approximation has been very useful in performing inference on latent Gaussian models, such as Gaussian processes [76]. The Integrated Nested Laplace Approximation (INLA) [77,78] can be used to compute the posteriors of the model parameters in the case of structured additive regression models where the predictor depends on a sum of functions of a set of covariates, and the number of hyperparameters is small (≤ 6). This is accomplished by setting up a grid of hyperparameter values where the posterior of the hyperparameter values given the data has been approximated using the Laplace approximation. Then the Laplace approximation is used to compute the marginal posteriors given the data and the hyperparameter values across the grid. The product of the hyperparameter posteriors (given the data) and the marginals (given the data and the hyperparameters) can then be numerically integrated over the hyperparameters to obtain the desired posterior marginals. Another method to approximate the marginals based on expectation propagation [79] has been proposed by Cseke and Heskes [80]. They demonstrated that this method is typically more accurate than INLA and works in cases where the Laplace approximation fails.

4.2. Importance sampling

Importance Sampling [81] allows one to find expectation values with respect to one distribution $p(x)$ by computing expectation values with respect to a second distribution $q(x)$ that is easier to sample from. The expectation value of $f(x)$ with respect to $p(x)$ is given by

$$\langle f(x) \rangle_p = \frac{\int f(x) p(x) dx}{\int p(x) dx}. \quad (31)$$

Note that one can write the distribution $p(x)$ as $\frac{p(x)}{q(x)} q(x)$ where the only theoretical requirement is that $q(x)$ must be non-zero wherever $p(x)$ is non-zero. This allows one to rewrite the expectation value above as

$$\langle f(x) \rangle_p = \frac{\int f(x) \frac{p(x)}{q(x)} q(x) dx}{\int \frac{p(x)}{q(x)} q(x) dx} \quad (32)$$

$$= \frac{\left\langle f(x) \frac{p(x)}{q(x)} \right\rangle_q}{\left\langle \frac{p(x)}{q(x)} \right\rangle_q}, \quad (33)$$

which can be approximated with samples from $q(x)$ by

$$\langle f(x) \rangle_p \approx \frac{\sum_{i=1}^N f(x_i) \frac{p(x_i)}{q(x_i)}}{\sum_{i=1}^N \frac{p(x_i)}{q(x_i)}}, \quad (34)$$

where the samples $x = x_1, x_2, \dots, x_N$ are drawn from $q(x)$. This works well as long as the ratio defined by $p(x)/q(x)$ does not attain extreme values. Importance sampling is a generally valid method useful even in cases where $q(x)$ is not Gaussian, as long as $q(x)$ is easier to sample from than $p(x)$ using techniques such as existing random number generators or MCMC.

Importance sampling can be used to compute ratios of evidence values in a similar fashion by writing [81]

$$\frac{Z_p}{Z_q} = \frac{\int p(x) dx}{\int q(x) dx} \quad (35)$$

which can be written as

$$\frac{Z_p}{Z_q} = \frac{\int \frac{p(x)}{q(x)} q(x) dx}{\int q(x) dx} \quad (36)$$

$$= \left\langle \frac{p(x)}{q(x)} \right\rangle_q \quad (37)$$

which can be approximated with samples from $q(x)$ by

$$\left\langle \frac{p(x)}{q(x)} \right\rangle_q \approx \frac{\sum_{i=1}^N \frac{p^2(x_i)}{q^2(x_i)}}{\sum_{i=1}^N \frac{p(x_i)}{q(x_i)}}. \quad (38)$$

However, again the function $p(x)$ must be close to $q(x)$ to avoid extreme ratios, which will cause problems for the numeric integration.

4.3. Analogy to statistical physics

Techniques for evaluating the evidence can build on numerical methods in statistical physics because there is a close analogy between both fields. A key quantity in equilibrium statistical mechanics is the canonical partition function

$$Z(\beta) = \int dx e^{-\beta E(x)} \quad (39)$$

where x are the configurational degrees of freedom of a system governed by the energy $E(x)$ and β is the inverse temperature. Because x is typically very high-dimensional, the partition function can only be evaluated numerically. Instead of computing $Z(\beta)$ directly by solving the high-dimensional integral (39), it is convenient to compute the *Density of States* (DOS)

$$g(E) = \int dx \delta(E - E(x)) \quad (40)$$

where δ is Dirac's delta function. The partition function and the DOS are linked via a Laplace transform

$$Z(\beta) = \int dE g(E) e^{-\beta E}. \quad (41)$$

Therefore, knowing either of the two functions suffices to characterize equilibrium properties of the system and compute, for example, free energies and heat capacities.

In a Bayesian application, the model parameters m play the role of the system's degrees of freedom and the negative log likelihood can be viewed as an energy function $E(m) = -\log P(d|m, M, I)$. For a given data set d , we write the DOS as

$$g(E) = \int dm P(m|M, I) \delta[E - E(m)]. \quad (42)$$

The evidence can then be written as a one-dimensional integral over the DOS:

$$\begin{aligned} P(d|M, I) &= \int dE g(E) e^{-E} \\ &= \int dm P(m|M, I) \int dE \delta[E + \log P(d|m, M, I)] e^{-E} \\ &= \int dm P(m|M, I) P(d|m, M, I). \end{aligned} \quad (43)$$

Therefore, knowledge of $g(E)$ allows us to compute the evidence in the same way as the canonical partition function (41) can be evaluated through a Laplace transform of the DOS [82].

Physics-inspired algorithms for evaluating the evidence aim to compute either the partition function $Z(\beta)$ at $\beta = 1$ or the density of states. The previous class of methods comprises path sampling [83], parallel tempering [84,85], annealed importance sampling [86] and other thermal methods that simulate a modified version of the posterior:

$$[P(d|m, M, I)]^\beta P(m|M, I) \quad (44)$$

where the likelihood has been raised to a fractional power. By letting β vary between zero and one, we can smoothly bridge between the prior and the posterior. A recent DOS-based algorithm called nested sampling [87,88,17,6] is discussed in Section 4.7.

4.4. Path sampling and thermodynamic integration

The method of *path sampling* is based on the calculation of free energy differences in thermodynamics [83]. The method focuses on the estimation of the difference between the logarithm of two distributions p_0 and p_1 , which depend on model parameters. One can connect the two distributions by a “path” through a space of distributions by defining what is called the geometric path

$$p(x|\beta) \propto p_0(x)^{1-\beta} p_1(x)^\beta \quad (45)$$

where the parameter β can vary freely from $\beta = 0$ to $\beta = 1$ so that at the endpoints we have that $p(x|\beta = 0) = p_0(x)$ and $p(x|\beta = 1) = p_1(x)$. By letting $E = \log[p_0/p_1]$ we can establish a direct relation with the canonical ensemble; the normalizing constant is the partition function:

$$\begin{aligned} Z(\beta) &= \int dx p_0(x)^{1-\beta} p_1(x)^\beta \\ &= \int dx p_0(x) e^{-\beta E(x)}. \end{aligned} \quad (46)$$

The log partition function can be estimated using samples from $p(x|\beta)$ in the following way. We have

$$\begin{aligned} \partial_\beta \log Z(\beta) &= -\frac{1}{Z(\beta)} \int dx E(x) p_0(x) e^{-\beta E(x)} \\ &= \langle \log[p_1/p_0] \rangle_\beta \end{aligned} \quad (47)$$

where $\langle \cdot \rangle_\beta$ denotes the expectation with respect to the bridging distribution $p(x|\beta)$. Integration of the previous equation yields

$$\begin{aligned} \log[Z(1)/Z(0)] &= \int_0^1 d\beta \partial_\beta \log Z(\beta) \\ &= \int_0^1 d\beta \langle \log[p_1/p_0] \rangle_\beta. \end{aligned} \quad (48)$$

By choosing a finely spaced β -path we can approximate the ratio of the normalization constants $Z(1)/Z(0)$ by a sum over the expected energy $\log[p_0/p_1]$ (log likelihood ratio) over each of the bridging distributions:

$$\log[Z(1)/Z(0)] \approx \sum_i \langle \log[p_1/p_0] \rangle_{\beta_i} (\beta_{i+1} - \beta_i). \quad (49)$$

This approach is called *thermodynamic integration*. It is also possible to estimate the DOS from samples produced along a thermal path bridging between the prior and posterior and thereby obtain an alternative estimate of the evidence that is sometimes more accurate than thermodynamic integration [82,89].

If we choose

$$p_0(m) = P(m|M, I) \text{ and } p_1(m) = P(m|M, I) P(d|m, M, I),$$

we can use path sampling in combination with thermodynamic integration to obtain the log-evidence because

$$Z(0) = \int dm p_0(m) = 1 \text{ and } Z(1) = \int dm p_1(m) = P(d|M, I).$$

In case we want to compare two models M_1, M_2 that share the same parameters m , we can use thermodynamic integration to estimate the log odds ratio (11) by defining

$$p_{i-1}(m) = P(m|M_i, I) P(d|m, M_i, I) \quad (i = 1, 2)$$

and sampling from the following family of bridging distributions

$$\begin{aligned} p(x|\beta) &\propto [P(m|M_1, I) P(d|m, M_1, I)]^{1-\beta} \\ &\quad \times [P(m|M_2, I) P(d|m, M_2, I)]^\beta \end{aligned} \quad (50)$$

For the special case that both models also share the same prior, $P(m|M_1, I) = P(m|M_2, I) = P(m|I)$, this simplifies to

$$p(m|\beta) \propto P(m|I) [P(d|m, M_1, I)]^{1-\beta} [P(d|m, M_2, I)]^\beta. \quad (51)$$

By drawing models from the mixed posterior $p(m|\beta)$ the log odds ratio can be computed directly using thermodynamic integration. An open problem relevant to all thermal methods using a geometric path (45) is where to place the intermediate distributions. This becomes increasingly difficult for complex systems that show a phase transition.

Ensemble Annealing [90], a variant of *simulated annealing* [91], aims to circumvent this problem by constructing an optimal temperature schedule in the course of the simulation. This is achieved

by controlling the relative entropy between successive intermediate distributions: After simulating the system at a current temperature, the new temperature is chosen such that the estimated relative entropy between the current and the new distribution is constant. Ensemble annealing can be viewed as a generalization of nested sampling (see Section 4.7) to general families of bridging distributions such as the geometric path (45). Ensemble annealing has been applied to various systems showing first- and second-order phase transitions such as Ising, Potts, and protein models [90].

4.5. Annealed importance sampling

Annealed Importance Sampling (AIS) [86] is closely related to other annealing methods such as simulated annealing but does not rely on thermodynamic integration. AIS generates multiple independent sequences of states $\{x_0^{(j)}, x_1^{(j)}, \dots, x_i^{(j)}, \dots\}$ where $x_i^{(j)}$ is a sample from the i -th intermediate distribution p_i bridging between the initial distribution p_0 and the destination distribution p_1 . For example, in case we are using the geometric bridge (45) the states follow

$$x_i^{(j)} \sim p_0(x)^{1-\beta_i} p_1(x)^{\beta_i} \quad (52)$$

where the superscript j enumerates the sequences. Sampling of $x_i^{(j)}$ is typically achieved by starting a Markov chain sampler from the precursor state $x_{i-1}^{(j)}$. Each of the generated sequences is assigned an importance weight

$$w^{(j)} = \prod_i \frac{p_{i+1}(x_i^{(j)})}{p_i(x_i^{(j)})}. \quad (53)$$

Neal has shown [86] that the average of the importance weights is an unbiased estimator of the ratio of the normalizing constants:

$$\frac{Z(1)}{Z(0)} \approx \frac{1}{M} \sum_{j=1}^M w^{(j)} = \frac{1}{M} \sum_{j=1}^M \prod_i \frac{p_{i+1}(x_i^{(j)})}{p_i(x_i^{(j)})}. \quad (54)$$

It is important to note that the annealing sequence is simulated multiple times, and that the partition function is obtained from the importance weights $w^{(j)}$ by an arithmetic average rather than a geometric average. For the special case of the geometric bridge, the AIS estimator is

$$\frac{Z(1)}{Z(0)} \approx \frac{1}{M} \sum_j \exp \left\{ \sum_i (\beta_{i+1} - \beta_i) \log [p_1(x_i^{(j)})/p_0(x_i^{(j)})] \right\}. \quad (55)$$

On the other hand, if we apply thermodynamic integration [Eq. (49)] to the sequences sampled during AIS, we obtain

$$\log[Z(1)/Z(0)] \approx \sum_i (\beta_{i+1} - \beta_i) \frac{1}{M} \sum_j \log [p_1(x_i^{(j)})/p_0(x_i^{(j)})]. \quad (56)$$

Both estimators are closely related but not identical. To see this, let us rewrite the estimate obtained by thermodynamic integration:

$$\frac{Z(1)}{Z(0)} \approx \exp \left\{ \frac{1}{M} \sum_j \sum_i (\beta_{i+1} - \beta_i) \log [p_1(x_i^{(j)})/p_0(x_i^{(j)})] \right\} \quad (57)$$

$$\approx \exp \left\{ \frac{1}{M} \sum_j \log w^{(j)} \right\} = \left(\prod_j w^{(j)} \right)^{1/M}. \quad (58)$$

This shows that AIS estimates the ratio of partition functions by an arithmetic average over the importance weights, whereas thermodynamic integration averages the importance weights $w^{(j)}$ geometrically. Neal's analysis as well as results from non-equilibrium thermodynamics (e.g. [92]) show that the AIS estimator is valid even if the sequences of states are not in equilibrium.

4.6. Variational Bayes

Another technique called *Ensemble Learning* [93–95], or *Variational Bayes* (VB) [96–100,74,101], is named after Feynman's variational free energy method in statistical mechanics [102]. As such, it is yet another example of how methods developed in thermodynamics and statistical mechanics have had an impact in machine learning and inference.

We consider a normalized probability density $Q(m)$ on the set of model parameters m , such that

$$\int dm Q(m) = 1. \quad (59)$$

While not obviously useful, the log-evidence can be written as

$$\log P(M|I) = \int dm Q(m) \log P(M|I). \quad (60)$$

Using the product rule, this can be written as

$$\log P(M|I) = \int dm Q(m) \log \frac{P(M, m|I)}{P(m|M, I)} \quad (61)$$

$$= \int dm Q(m) \log \left[\frac{P(M, m|I) Q(m)}{P(m|M, I) Q(m)} \right]. \quad (62)$$

This expression can be broken up into the sum of the negative free energy

$$F(Q(m), P(M, m|I)) = \int dm Q(m) \log \frac{P(M, m|I)}{Q(m)} \quad (63)$$

and the Kullback–Leibler (KL) divergence

$$KL[Q(m) \| P(m|M, I)] = \int dm Q(m) \log \frac{Q(m)}{P(m|M, I)} \quad (64)$$

by

$$\log P(M|I) = F(Q(m), P(M, m|I)) + KL[Q(m) \| P(m|M, I)], \quad (65)$$

which is the critical concept behind variational Bayes.

The properties of the KL divergence expose an important relationship between the negative free energy and the evidence. First, the KL divergence is zero when the density $Q(m)$ is equal to the posterior $Q(m) = P(m|M, I)$. For this reason, $Q(m)$ is referred to as the approximate posterior. Furthermore, since the KL divergence is always non-negative, we have that

$$\log P(M|I) = \max_Q F(Q(m), P(M, m|I)) \geq F(Q(m), P(M, m|I)) \quad (66)$$

so that the negative free energy is a lower bound to the log-evidence.

The main idea is to vary the density $Q(m)$ (approximate posterior) so that it approaches the posterior $P(m|M, I)$. One cannot do this directly through the KL divergence since the evidence, which is the normalization factor for the posterior, is not known. Instead, by maximizing the negative free energy in (65), which is the same as minimizing the free energy, the negative free energy approaches the log-evidence and the approximate posterior $Q(m)$ approaches the posterior. However, this presents a technical difficulty in that the integral for the negative free energy (63) will not be analytically solvable for arbitrary $Q(m)$. The approach generally taken involves a concept from the mean field approximation in statistical mechanics [103] where a non-factorizable function is replaced by one that is factorizable

$$Q(m) = Q(m_0) Q(m_1) \quad (67)$$

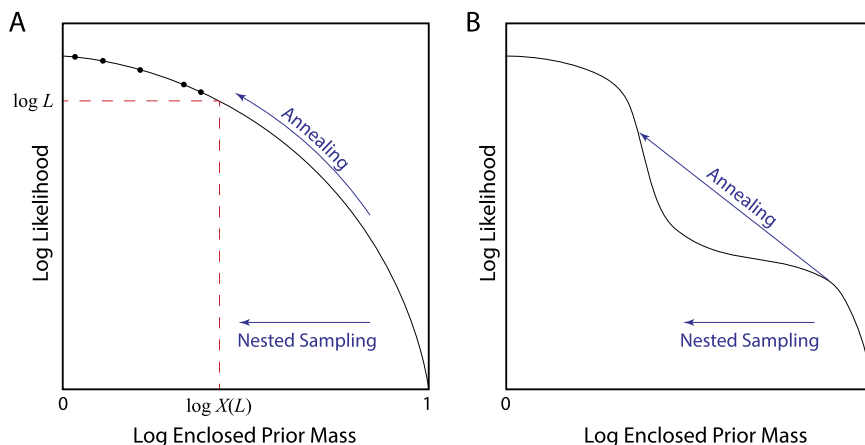


Fig. 1. A: An illustration of the amount of log prior mass $\log X(L)$ with a log likelihood greater than $\log L$. Nested sampling relies on forming a nested set of likelihood boundaries within which the N walkers are uniformly distributed. As such, nested sampling contracts the prior volume to higher likelihood at a steady rate based on $\log X$ [17]. On the other hand, tempering methods, such as simulated annealing, advance on high likelihood regions by following the slope $\frac{d \log L}{d \log X}$ of the curve. B: Tempering methods, which slowly turn on the likelihood L^β with the inverse temperature parameter β , are designed to follow the concave hull of the log likelihood. In situations where the slope becomes convex, one must jump from one phase (local maximum in evidence mass) to another, which is why tempering methods typically fail at phase transitions. Nested sampling, which contracts the prior volume, is not hampered by such features in the log likelihood curve as a function of prior mass.

where the set of model parameters m can be divided into two disjoint sets m_0 and m_1 so that $m_0 \cap m_1 = \emptyset$ and $m_0 \cup m_1 = m$.

The negative free energy (63) can then be written as [74]

$$\begin{aligned} F &= \int dm Q(m) \log \frac{P(M, m|I)}{Q(m)} \\ &= \int \int dm_0 dm_1 Q(m_0) Q(m_1) \log \frac{P(M, m_0, m_1|I)}{Q(m_0) Q(m_1)} \\ &= \int dm_0 Q(m_0) \left[\int dm_1 Q(m_1) \log P(M, m_0, m_1|I) \right] \\ &\quad - \int dm_0 Q(m_0) \log Q(m_0) + C \\ &= \int dm_0 Q(m_0) I(m_0) - \int dm_0 Q(m_0) \log Q(m_0) + C \end{aligned}$$

where the constant C consists of terms that do not depend on $Q(m_0)$ and

$$I(m_0) = \int dm_1 Q(m_1) \log P(M, m_0, m_1|I). \quad (68)$$

The negative free energy can then be expressed in terms of a KL-divergence by writing $I(m_0) = \log(\exp(I(m_0)))$

$$F = KL[Q(m_0) \parallel \exp(I(m_0))] + C, \quad (69)$$

which is minimized when

$$Q(m_0) \propto \exp(I(m_0)). \quad (70)$$

This implies that not only can the posterior be approximated with $Q(m)$, but also the analytic form of the component posteriors can be determined. This is known as the *free-form approximation* [74], which applies, in general, to the conjugate exponential family of distributions [97,104,105], and can be extended to non-conjugate distributions [97,106].

Since the negative free energy (69) is a lower bound to the log-evidence, the log-evidence can be estimated by minimizing the negative free energy, so that the approximate posterior $Q(m)$ approaches the posterior.

4.7. Nested sampling

Nested sampling [87,88,17,6] relies on stochastic integration to numerically compute the evidence of the posterior probability. In

contrast to the thermal algorithms discussed so far, nested sampling aims to estimate the DOS or rather its cumulative distribution function

$$\begin{aligned} X(L) &= \int_{-\infty}^{-\log L} dE g(E) \\ &= \int_{P(d|m, M, I) > L} dm P(m|M, I) \end{aligned} \quad (71)$$

which calculates the prior mass $X \in [0, 1]$ contained in the likelihood contour $P(d|m, M, I) > L \equiv e^{-E}$. We can now write the evidence integral as

$$\begin{aligned} Z &= \int_{-\infty}^{\infty} dE g(E) e^{-E} \\ &= \int_0^1 dX L(X) \\ &\approx \sum_i L_i (X_{i-1} - X_i) \end{aligned} \quad (72)$$

where the likelihood $L(X)$ is understood as a function of the cumulative DOS or prior mass (71). Because $L(X)$ is unknown for general inference problems, we have to estimate it. Nested sampling does this by estimating its inverse function $X(L)$ using N walkers that explore the prior constrained by a lower/upper bound on the likelihood/energy (Fig. 1A). Since X decreases monotonically in likelihood, we can sort the unknown prior masses associated with each walker by sorting them according to likelihood. The walker with worst likelihood will enclose the largest prior mass. The maximum mass can be estimated using order statistics:

$$X_{\max} \sim N \frac{X_{\max}^{N-1}}{X(L)} \quad (73)$$

where the walkers have been numbered such that they increase in likelihood $L_1 < L_2 < \dots < L_N$ and thus $X_{\max} \equiv X_1 > X_2 > \dots > X_N$. The worst likelihood L_1 will define the lower likelihood bound in the next iteration. Walkers 2 to N will, by construction, already attain states that are also valid samples from the prior truncated

at L_1 such that we only have to replace the first walker. This can be done by randomly selecting one among the $N - 1$ surviving states and evolving it within the new contour L_1 using a Monte Carlo procedure. The initial states are obtained by sampling from the prior (i.e. the lower bound on the likelihood is zero); the associated mass is $X(0) = 1$.

Rather than increasingly giving the data more and more weight as is done in thermal approaches, nested sampling focuses on states with high posterior probability by constructing a sequence of nested priors restricted to higher and higher likelihood regions. Thereby nested sampling locates the relevant states that contribute strongly to the evidence integral and simultaneously constructs an optimal sequence of likelihood contours. In path sampling, the geometric path must be typically chosen by the user; nested sampling elegantly circumvents this problem.

Another advantage of nested sampling is that each likelihood bound in the nested sequence compresses the prior volume by approximately the same factor, which allows nested sampling to handle first order phase transitions. In contrast, tempering methods such as simulated annealing and parallel tempering compress based on steps in temperature ($\propto L^{1/T}$), and as a result they typically fail at phase transitions (Fig. 1B).

Often a practical difficulty of applying nested sampling is the requirement to sample from the prior subject to a hard constraint on the likelihood. Mukherjee et al. [107] developed a version of nested sampling that fits an enlarged ellipse around the walkers and samples uniformly from that ellipse until a sample is drawn that has a likelihood exceeding the old minimum. Sampling within the hard constraint is also made difficult when the distribution is multi-modal. MultiNest [108,109,30] was developed to handle multi-modal distributions by using K-means clustering to cluster the walkers into a set of ellipsoids. At each iteration, MultiNest replaces the walker with the worst likelihood by a new walker generated by randomly selecting an ellipsoid (uniformly) and sampling uniformly from within the bounds of that ellipsoid. These ellipsoids serve to allow one to detect and characterize multiple peaks in the distribution. However, the method has two drawbacks in which accurate K-means clustering limits the dimensionality of the problem to tens of parameters, and the elliptical regions may not always cover the high likelihood regions of the parameter space.

Other variants of nested sampling couple the technique with Hamilton Monte Carlo [110] or Galilean Monte Carlo [111,112,20], which sample within the hard likelihood constraint by considering the step size to be determined by some particle dynamics depending on the particle velocity, and using that velocity and likelihood gradient to reflect off of the hard likelihood boundary. This has been demonstrated to result in improved exploration in cases of multi-modal distributions and distributions with curved degeneracies.

Another possible way to facilitate sampling from within the hard likelihood constraint is to introduce additional “demon” variables that smooth the constraint boundary and push the walkers away from it [113]. This approach can help to solve complex inference problems as they arise, for example, in protein structure determination, at the expense of introducing additional algorithmic parameters.

Diffusive Nested Sampling [114] is a variant of nested sampling that monitors the log likelihood values during the MCMC steps and creates nested levels such that each level covers approximately e^{-1} of the prior mass of the previous level. This allows the relative enclosed prior mass of the nested levels to be estimated more accurately than in nested sampling. Samples are then obtained from a weighted mixture of the current level and the previous levels so that a mixture of levels is diffusively explored facilitating travel be-

tween isolated modes and allowing a more refined estimate of the log evidence.

5. Practical examples

In this section we consider a set of four practical examples where the Bayesian evidence is both calculated and used in different ways. The purpose of this section is not to compare one computational method against another, since given the large number of techniques available, this would require a more extensive treatment. Instead, the goal is to demonstrate the utility of Bayesian model selection in several examples both relevant to signal processing and spanning the domain sciences.

The first example focuses on the problem of signal detection where the evidence, which is computed analytically, is used to test between two models: signal present and signal absent. The second example focuses on using the evidence, estimated numerically by nested sampling, to select the model order of a Gaussian mixture model of the spatial sensitivity function of a light sensor. The third example relies on the application of the evidence, estimated using MultiNest, to select among a set of exoplanet models each exhibiting different combinations of photometric effects. The final example selects a molecular mechanics force field approximately describing atomic interactions in proteins by computing the evidence of nuclear magnetic resonance (NMR) data.

5.1. Signal detection

In this example, based on the work by Mubeen and Knuth [115], we consider a practical signal detection problem where the log odds-ratio can be analytically derived. The result is a novel signal detection filter that outperforms correlation-based detection methods in the case where both the noise variance and the variance in the overall signal amplitude is known. While this detection filter was originally designed to be used in brain-computer interface (BCI) applications, it is applicable to signal detection in general (with slight modification).

We consider the problem of detecting a stereotypic signal, $s(t)$, which is modeled by a time-series with T time points. This signal has the potential to be recorded from M detector channels with various (potentially negative) coupling weights C_m where the index m refers to the m th channel. Last, and perhaps more specific to the BCI problem, we consider that the overall amplitude of the emitted signal waveshape $s(t)$ can vary. This is modeled using a positive-valued amplitude parameter α , which is the only free parameter as it is assumed that the coupling weights C_m and the signal waveshape $s(t)$ are known.

There are two states to be considered: *signal absent* (null hypothesis) and *signal present*. We model the signal absent state as noise only

$$M_N : x_m(t) = n_m(t) \quad (74)$$

where M_N denotes the “noise-only” model, $x_m(t)$ denotes the signal time-series recorded in the m th channel and $n_m(t)$ refers to the noise signal associated with the m th channel. The signal present state is modeled as signal plus noise by

$$M_{S+N} : x_m(t) = \alpha C_m s(t) + n_m(t) \quad (75)$$

where the symbol M_{S+N} denotes the “signal-plus-noise” model and α is the amplitude of the signal $s(t)$, which is coupled to each of the m detectors with weights C_m .

The *odds-ratio* can be written as the ratio of evidences

$$\text{OR} = \frac{P(X|M_{S+N}, I)}{P(X|M_N, I)} \equiv \frac{Z_{S+N}}{Z_N} \quad (76)$$

where X represents the available data, which here will be the recorded time series vector $\mathbf{x}(t) = \{x_1(t), x_2(t), \dots, x_M(t)\}$, and I represents any relevant prior information including the coupling weights C_m and the signal waveshape $s(t)$. The two evidence values can be written as

$$Z_N = P(X|M_N, I) \quad (77)$$

$$= P(\mathbf{x}(t)|\mathbf{n}(t), I) \quad (78)$$

and

$$Z_{S+N} = P(X|M_{S+N}, I) \quad (79)$$

$$= \int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha P(\alpha|I)P(\mathbf{x}(t)|\mathbf{n}(t), I) \quad (80)$$

where the latter is marginalized over the amplitude range $[\alpha_{\min}, \alpha_{\max}]$ of the signal α since we only care to detect the signal. Here $P(\alpha|I)$ represents the prior probability for the amplitude parameter α . Note also that $\mathbf{x}(t)$ and $\mathbf{n}(t)$ without subscripts refer to the vector of time series over each of the detector channels.

Assuming that the noise signals $\mathbf{n}(t)$ have identical characteristics in each channel, we assign a Gaussian likelihood with a standard deviation of σ_n to both models. Note that this is not quite the same as assuming that the signals are Gaussian distributed, but rather this is the maximum entropy assignment where both the mean and squared deviation from the mean are known to be relevant quantities. For the “noise-only” model there are no model parameters and the likelihood is equal to the evidence (78)

$$Z_N = (2\pi\sigma_n^2)^{-MT/2} \exp\left[-\frac{1}{2\sigma_n^2} \sum_{m=1}^M \sum_{t=1}^T x_m^2(t)\right]. \quad (81)$$

In the “signal-plus-noise” model we have the Gaussian likelihood

$$P(\mathbf{x}(t)|\alpha, \mathbf{n}(t), I) = (2\pi\sigma_n^2)^{-MT/2} \exp\left[-\frac{1}{2\sigma_n^2} \sum_{m=1}^M \sum_{t=1}^T (x_m(t) - \alpha C_m s(t))^2\right]. \quad (82)$$

By assigning a (potentially-truncated) Gaussian prior to the amplitude parameter α ,

$$P(\alpha|I) = \frac{1}{Z_\alpha} \exp\left[-\frac{1}{2\sigma_\alpha^2} (\alpha - \hat{\alpha})^2\right], \quad (83)$$

with a normalization constant Z_α given by

$$Z_\alpha = \int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha \exp\left[-\frac{1}{2\sigma_\alpha^2} (\alpha - \hat{\alpha})^2\right], \quad (84)$$

one can integrate the likelihood (82) to find the evidence of the “signal-plus-noise” model (80).

By defining

$$D = S^2 + \sum_{m=1}^M \sum_{t=1}^T C_m^2 s^2(t) \quad (85)$$

$$E = S^2 \hat{\alpha} + \sum_{m=1}^M \sum_{t=1}^T C_m x_m(t) s(t) \quad (86)$$

$$F = S^2 \hat{\alpha}^2 + \sum_{m=1}^M \sum_{t=1}^T x_m^2(t), \quad (87)$$

where

$$S^2 = \frac{\sigma_n^2}{\sigma_\alpha^2}, \quad (88)$$

we can complete the square in the exponent and write the odds ratio as

$$\frac{Z_{S+N}}{Z_N} = \frac{\int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha P(\alpha|I)P(\mathbf{x}(t)|\mathbf{n}(t), I)}{Z_N} \quad (89)$$

$$= \exp\left[-\frac{1}{2\sigma_n^2} (S^2 \hat{\alpha}^2 - E^2/D)\right] \frac{Z_d}{Z_\alpha} \quad (90)$$

where

$$Z_d = \int_{\alpha_{\min}}^{\alpha_{\max}} d\alpha \exp\left[-\frac{D}{2\sigma_n^2} (\alpha - E/D)^2\right]. \quad (91)$$

In general, these Gaussian integrals result in solutions involving the error function (erf) [116]

$$\int_a^b dx e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \frac{\sqrt{2\pi}\sigma^2}{2} \left[\operatorname{erf}\left(\frac{b-\mu}{\sqrt{2}\sigma}\right) + \operatorname{erf}\left(\frac{\mu-a}{\sqrt{2}\sigma}\right) \right]. \quad (92)$$

If we restrict the signal amplitude to being positive, we have that $\alpha_{\min} = 0$ and $\alpha_{\max} = +\infty$ and the integrals (91) and (84) become

$$Z_d = \frac{\sqrt{2\pi}\sigma_n^2/D}{2} \left[1 + \operatorname{erf}\left(\frac{E}{\sqrt{2D}\sigma_n}\right) \right] \quad (93)$$

and

$$Z_\alpha = \frac{\sqrt{2\pi}\sigma_\alpha^2}{2} \left[1 + \operatorname{erf}\left(\frac{\hat{\alpha}}{\sqrt{2}\sigma_\alpha}\right) \right] \quad (94)$$

resulting in the log odds ratio

$$\log \text{OR}_+ = \frac{1}{2} \left[\left(\frac{E^2}{D\sigma_n^2} - \frac{\hat{\alpha}^2}{\sigma_\alpha^2} \right) + \log\left(\frac{S^2}{D}\right) \right] + \log\left(\frac{1 + \operatorname{erf}\left(\frac{E}{\sqrt{2D}\sigma_n}\right)}{1 + \operatorname{erf}\left(\frac{\hat{\alpha}}{\sqrt{2}\sigma_\alpha}\right)}\right), \quad (95)$$

where the subscript + indicates that the signal amplitude α is assumed to be positive.

However, if we consider allowing α to vary over the entire real line by setting $\alpha_{\min} = -\infty$ and $\alpha_{\max} = +\infty$ we find that

$$Z_d = \sqrt{2\pi}\sigma_n^2/D \quad (96)$$

and

$$Z_\alpha = \sqrt{2\pi}\sigma_\alpha^2, \quad (97)$$

which gives a simpler log odds ratio which lacks the term with the erf functions

$$\log \text{OR}_\pm = \frac{1}{2} \left[\left(\frac{E^2}{D\sigma_n^2} - \frac{\hat{\alpha}^2}{\sigma_\alpha^2} \right) + \log\left(\frac{S^2}{D}\right) \right], \quad (98)$$

where the subscript \pm indicates that the signal amplitude α ranges from $-\infty$ to ∞ .

The expression E (86) contains the cross-correlation term, which is what is typically used for the detection of a target signal in ongoing recordings. The log OR detection filters incorporate

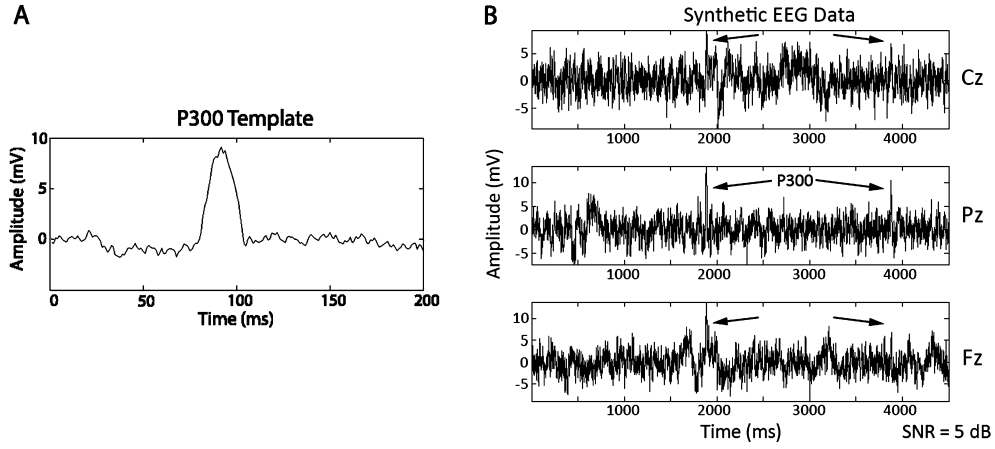


Fig. 2. A: The P300 template target signal. B: An example of three channels (Cz, Pz, Fz) of synthetic ongoing EEG with two P300 target signal events (indicated by the arrows) at an SNR of 5 dB.

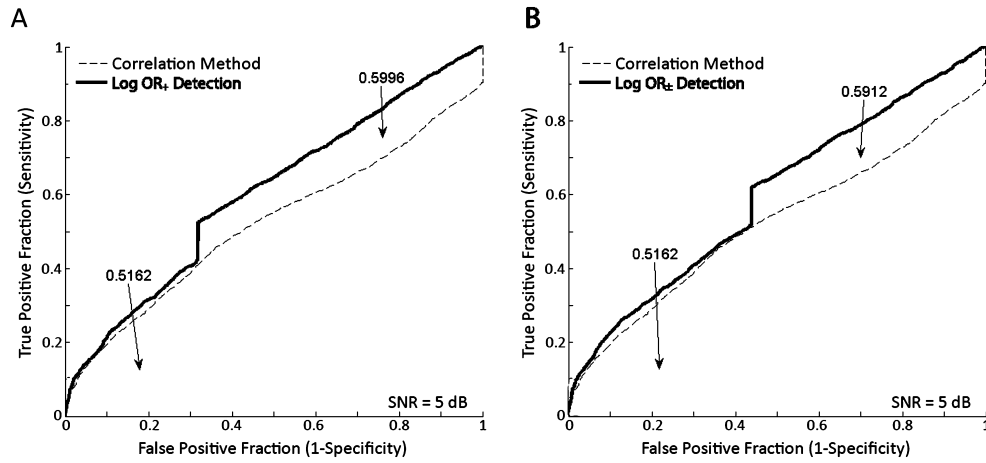


Fig. 3. A: This illustrates the ROC curves for both the Correlation Detection Method and the $\log OR_+$ Detection Method in the case of SNR = 5 dB. Note that the $\log OR_+$ Detection has a greater area under the curve (0.5996 as opposed to 0.5162 for Correlation), which indicates better performance over the Correlation Method. B: This figure illustrates the ROC curves for both the Correlation Detection Method and the $\log OR_{\pm}$ Detection Method in the case of SNR = 5 dB. While the $\log OR_{\pm}$ Detection performs better than Correlation (0.5912 as opposed to 0.5162 for Correlation), it does not do quite as well as $\log OR_+$ Detection in the case of SNR = 5 dB.

more information that leads to extra terms, which serve to aid in target signal detection.

Since the “signal-plus-noise” model (75) reduces to the “noise-only” model (74) as $\alpha \rightarrow 0$, one would expect that the odds ratio should go to one, $OR \rightarrow 1$, as $\alpha \rightarrow 0$. This can be accomplished by setting $\hat{\alpha} = 0$ and letting $\sigma_{\alpha} \rightarrow 0$ in which case the truncated Gaussian prior for α (83) collapses to a delta function. The odds ratio in (90) shows this limiting behavior as the argument of the exponential function approaches zero, and $Z_d/Z_{\alpha} \rightarrow 1$. Another way to ignore the signal is to set $C_m = 0$, in which case $D = S^2$, $E = S^2\hat{\alpha}$. Again the argument of the exponential function in (90) vanishes and ratio of the normalizing constants approaches one.

To analyze the performance of the log OR filters, we generated synthetic electroencephalographic (EEG) data representing both the EEG background and the P300 evoked response, which is a brain response commonly used in BCI applications [117] (Fig. 2A). Using the MATLAB code provided by Yeung, Bogacz, et al. [118], three channels of synthetic EEG data were generated to simulate recordings from scalp locations: Cz, Pz and Fz. A current dipole model was used to scale the synthetic recordings from the different channels [119]. The data from each of these channels consisted of 300 epochs each being 800 ms in length and comprised of 200 samples, which is consistent with a sampling rate of 250 Hz. Thirty epochs were selected to each host a single stereotypic P300 re-

sponse at random latencies. The remaining 270 epochs exhibited only ongoing background EEG (noise).

To study the effect of the Signal-to-Noise-Ratio (SNR) on the log OR filter performance, we created 17 data sets where the SNR, calculated by the formula

$$SNR_{dB} = 10 \log_{10} \left(\frac{A_{signal}}{A_{noise}} \right)^2, \quad (99)$$

was varied in integral steps from -6 dB to 7 dB as well as 10, 15 and 20 dB covering the typical SNR range seen in BCI and EEG applications. Fig. 2B illustrates synthetic ongoing EEG recordings with two target P300 signals (Fig. 2A) at an SNR of 5 dB.

The selection of a detection threshold value is a difficult task. As the detection threshold increases, the sensitivity decreases while the specificity increases, which means that the false positive fraction (1-specificity) decreases. To study the performance of the log OR detection filter we compared it to the standard Correlation Method by producing Receiver Operating Characteristics (ROC) curves. To do this we calculate sensitivity and (1-specificity) for each distinct value of the detection measure (i.e. log OR/Correlation) to consider it as a candidate for detection cutoff. By plotting (1-specificity) versus sensitivity, the efficacy of the detection method can be quantified by the area under the ROC curve [120]. Fig. 3 compares the ROC curves of the Correlation Method with the $\log OR_+$ Method (Fig. 3A) and the $\log OR_{\pm}$ Method (Fig. 3B) ob-

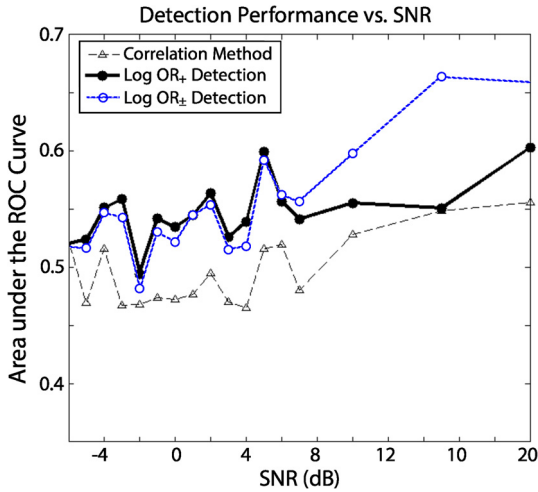


Fig. 4. A comparison of the performance of the two log OR methods, logOR₊ and logOR_±, and the Correlation Method as quantified by the areas under their respective ROC curves as a function of SNR. The logOR₊ filter consistently performs better for low SNR, but is outperformed by logOR_± Detection for SNR > 5 dB.

tained for target signals with an SNR of 5 dB. These figures indicate that at this particular SNR, the log OR Detection filters outperform the traditional Correlation Method. Fig. 4 provides a comparison of the performance of the two log OR methods, logOR₊ and logOR_±, and the Correlation Method as quantified by the areas under their respective ROC curves as a function of SNR for SNRs ranging from −6 dB to 20 dB. The results indicate that the log OR detection filters based on Bayesian model testing consistently outperformed traditional Correlation. Moreover, we see that the logOR₊ filter consistently performs better for low SNR, but is outperformed by logOR_± Detection for SNR > 5 dB.

5.2. Light sensor characterization

In this example, based on the work by Malakar, Gladkov and Knuth [121], we demonstrate the use of Bayesian evidence to select the model order for a Gaussian mixture model of a light sensor, which was used in a robotics application [121]. The problem involved identifying an accurate and efficient model of a LEGO light sensor (LEGO #9844). The sensor consists of a photodiode-LED pair where the LED is used to illuminate the surface and the photodiode is used to measure the intensity of the reflected light. The sensor integrates the light arriving from a spatially distributed region within its field of view, weighted by its spatial sensitivity function (SSF). The goal was to model the SSF so that we could make accurate predictions of how the light sensor would respond when placed above a surface with a known albedo pattern. We considered a mixture of Gaussians (MoG) model for the SSF in the sensor frame $(x', y') = (x - x_i, y - y_i)$,

$$SSF(x', y') = \frac{1}{K} \sum_{n=1}^N a_n \exp\left\{A_n(x' - u'_n)^2 + B_n(y' - v'_n)^2 + 2C_n(x' - u'_n)(y' - v'_n)\right\} \quad (100)$$

where a_n and (u'_n, v'_n) denote the amplitude and center of the n th Gaussian, respectively, where its covariance matrix elements are denoted by A_n , B_n and C_n . The factor K is the normalizing constant to ensure that the SSF integrates to unity in the case of a white surface.

The MoG model is sufficiently general to be able to well-describe the SSF by varying the number of Gaussians. We considered four models consisting of one, two, three and four Gaussians. Each Gaussian in the mixture requires six parameters to be es-

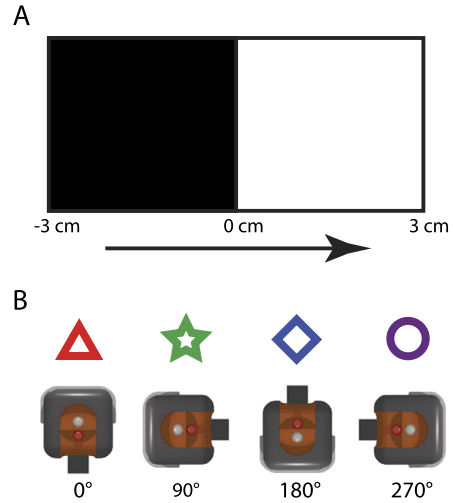


Fig. 5. A: An illustration of the black-and-white calibration surface. B: An illustration of the four orientations of the light sensor along with the colored symbols used to represent the intensities in Fig. 7 recorded at each orientation with respect to the calibration surface as indicated by the black arrows.

timated $\theta_n = a_n, u_n, v_n, A_n, B_n, C_n$, where the subscript n indexes the Gaussian in the mixture.

In order to infer the model, we collected data by performing a series of experiments by recording intensities as the sensor was moved along a known surface (Fig. 5A). The sensor was held at a height of 14 mm above the surface in one of four orientations illustrated in Fig. 5B. Intensities were recorded at increments of 1 mm steps as the sensor was moved in the direction of the arrow. In addition to the surface illustrated in Fig. 5A, we also presented the sensor with four corner patterns designed to break remaining symmetries.

Bayesian estimation of the MoG model parameters was performed using nested sampling with 300 samples, and was repeated 20 times for each model order to obtain uncertainty estimates of the log-evidence. We assigned uniform priors to the model parameters as well as a Student-t distribution to the likelihood. The nested sampling algorithm was iterated while monitoring the log-evidence, and was stopped when the change in the consecutive log-evidence values was less than $1e-8$.

Fig. 6 shows the four resulting MoG models of the SSF function. Table 1 shows the evidence values for the competing models. The 1-MoG model consisting of a single Gaussian had the greatest mean log-evidence of -665.1 , which is why it was selected as the optimal model. Fig. 7 compares the predictions (black) made by the 1-MoG SSF model to the observed intensities (red) showing excellent agreement. The resulting SSF model obtained by maximizing the log-evidence was found to be both accurate and efficient, and was selected for use in further studies involving that light sensor [121].

However, given the uncertainties of the log-evidence values of the models in Table 1, selection of a best model based on the log-evidence alone is not clear. These results suggest that this implementation of nested sampling is experiencing difficulties dealing with the multiple optima resulting from the degeneracies of the mixture model. This is similar to the challenge faced by Chib using Gibbs sampling [65], as noted, and solved, by Berkhof et al. [66]. In the Conclusion, we will make some additional comments on selecting an optimal model order based on log-evidence estimates.

5.3. Exoplanet detection

Our third example concerns the determination of the importance of various photometric effects in an exoplanetary system.

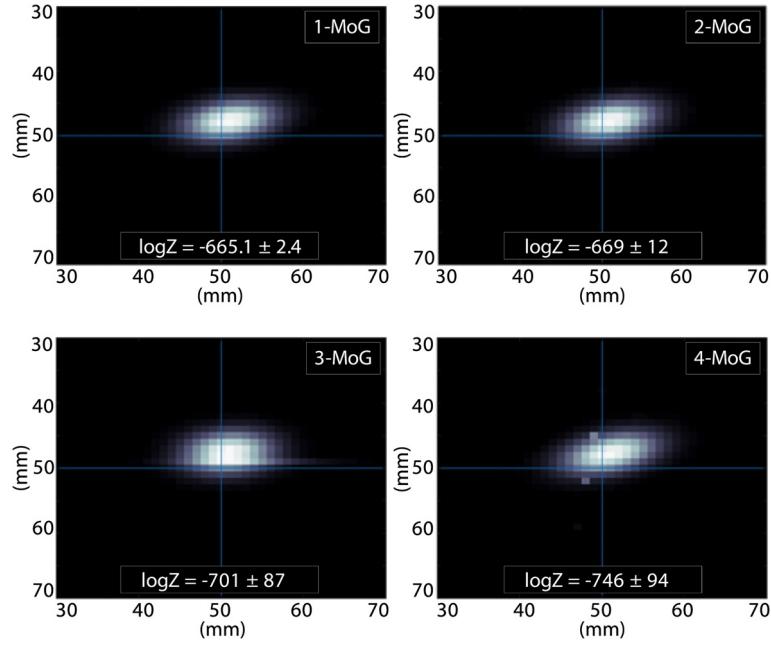


Fig. 6. An illustration of the four resulting MoG SSF Models along with their log-evidence values. The single Gaussian model (1-MoG) was found to have the greatest evidence, and therefore was selected as the optimal model of the light sensor SSF function.

Table 1

This table lists the log-evidence ($\log Z$) values estimated for the MoG SSF models of various model orders. The simplest model consisting of a single Gaussian (1-MoG) was found to be the most probable model. However, the increasing uncertainty in the log-evidence estimates strongly suggests that this implementation of nested sampling is experiencing difficulties in handling the degeneracies in the mixture model.

MoG model	$\log Z$	# of params
1 Gaussian	-665.1 ± 2.4	6
2 Gaussians	-669 ± 12	12
3 Gaussians	-701 ± 87	18
4 Gaussians	-746 ± 94	24

The details of this study by Placek, Knuth, et al. can be found in the following references [122,123]. Currently, the primary method of detecting and characterizing exoplanets involves the analysis of the time series resulting from the observations of unresolved light coming from a planetary system. The presence of exoplanets around distant stars is known to produce at least four physical mechanisms that affect the observed photometric signal in very specific ways. The first two effects originate from the planet itself. As the planet orbits its host star, it undergoes phases just as Venus and Mercury do in this Solar System from the perspective of Earth. This will cause photometric variations since the amount of reflected light off of the atmosphere or surface of the planet will change throughout the planet's orbit. By modeling the reflectance as Lambertian, one can model these stellar-normalized flux variations as

$$\frac{F_R(t)}{F_\star} = \frac{A_g R_p^2}{2 r(t)^2} (1 + \cos \theta(t)), \quad (101)$$

where A_g is the geometric albedo of the planet, which represents how effective the planet is at reflecting incident light back into space, R_p is the planetary radius, $r(t)$ is the planet-star separation distance, $\theta(t)$ is the angle between the observer's line-of-sight and the line connecting the star to the planet, and F_\star is the stellar flux. Similarly, planets have a temperature and therefore emit thermal radiation. This also contributes to the observed photometric signal and can be modeled for both day and night sides as

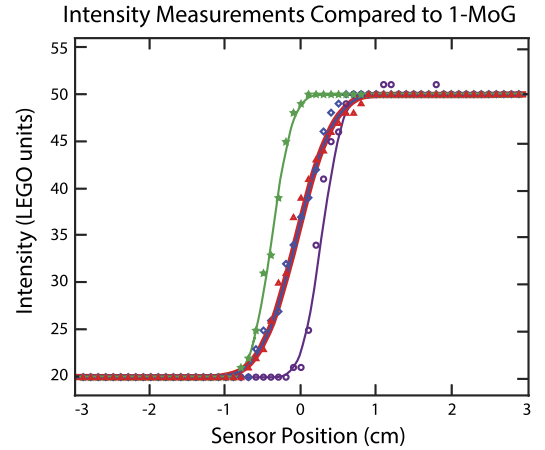


Fig. 7. The 1-MoG SSF model, with the maximum log-evidence, is used to predict the sensor intensity (solid curves) when applied to the black-and-white calibration surface and compare it with the recorded intensities (discrete symbols). The four symbols denote the four orientations of the sensor as indicated in Fig. 5. Note that since the 1-MoG SSF model is aligned with the measurement axes, the predicted curves for the 0° orientation (red triangles) and 180° orientation (blue diamonds) are identical and overlaid on top of one another in the center of the figure.

$$\frac{F_{T,d}(t)}{F_\star} = \frac{1}{2}(1 + \cos \theta(t)) \left(\frac{R_p}{R_\star} \right)^2 \frac{\int B(T_d)K(\lambda) d\lambda}{\int B(T_{eff})K(\lambda) d\lambda} \quad (102)$$

where R_\star and T_{eff} are the stellar radius and effective temperature, respectively, $B(T)$ is the spectral radiance of a blackbody, $T(d)$ is the day-time temperature of the planet, and $K(\lambda)$ is the instrument response as a function of wavelength λ . The expected stellar-normalized flux from the night-side $\frac{F_{T,n}(t)}{F_\star}$ is found using the night-side temperature of the planet T_n .

The remaining two effects are induced by the planet but involve the host star. Stars and planets both orbit the center of mass of the system. As the star revolves around the center of mass, an observer moving relative to that star will observe increases in the amount of flux emitted from the star as it approaches, and a decrease in flux as it recedes. This is known as Doppler beaming and is a rel-

ativistic effect. In the non-relativistic limit, the flux variations can be approximated as

$$\frac{F_B(t)}{F_\star} = 1 + 4\beta_r(t) \quad (103)$$

where $\beta_r(t)$ is the component of stellar velocity along the line-of-sight. This effect has the same frequency as the previous two, however the signal is shifted in phase by $\pi/2$. Finally, due to the proximity of the planet to the star, the planet will induce tides on the stellar surface causing the star to appear as a prolate spheroid. These tides will follow the planet in its orbit and result in flux variations at twice the orbital frequency since the cross-section of the star is changing throughout the orbit. This effect is approximated by

$$\frac{F_{\text{ellip}}(t)}{F_\star} = \beta \frac{M_p}{M_\star} \left(\frac{R_\star}{r(t)} \right)^3 \times [\cos^2(\omega + \nu(t)) + \sin^2(\omega + \nu(t)) \cos^2 i] \quad (104)$$

where β is the gravity darkening exponent, M_p and M_\star are the planetary and stellar masses, respectively, ω is the argument of periastron, $\nu(t)$ is the true anomaly, and i is the orbital inclination.

In order to obtain a predictive model for the total observed signal, one needs to sum the photometric contributions from each effect

$$F_{\text{pred}}(t) = F_\star \left(1 + \frac{F_p(t)}{F_\star} + \frac{F_{\text{boost}}(t)}{F_\star} + \frac{F_{\text{ellip}}(t)}{F_\star} + \frac{F_{\text{Th,d}}(t)}{F_\star} + \frac{F_{\text{Th,n}}(t)}{F_\star} \right). \quad (105)$$

Bayesian model selection allows one to effectively characterize exoplanetary systems. Each of these four effects can be present in the data to varying degrees, or completely absent. Thus, one can create a suite of models each comprised of a different subset of the four photometric effects. Since all four effects depend on the orbital orientation of the planet, model testing also allows one to test between circular and eccentric orbits. By calculating the evidence for each model, one could determine whether or not each effect is present in the data, and how large of a role each effect plays in describing the observed data.

As an example, we performed such model testing on data obtained from the Kepler Space Telescope for a confirmed exoplanet called KOI-13b. KOI-13b is known as a short-period hot Jupiter since it orbits its host star in just 1.7637 days and has a temperature of over 3500 K. This sort of exoplanet is expected to induce large ellipsoidal variations on its host star, produce significant thermal emission and less reflection. This is due to the fact that most of the reflective condensates in the atmosphere, such as water and ammonium, are essentially burned off, significantly decreasing the planetary albedo. A set of 18 models were applied to KOI-13b (shown in Table 2) and log-evidences were calculated for each one using the MultiNest algorithm [108,109,30], which is one of several inference engines included in our EXONEST Exoplanetary Explorer software suite [123].

In general, the noise is expected to be Gaussian-distributed about the mean exoplanetary signal. Therefore, a Gaussian log-likelihood of the form

$$\log L = -\frac{1}{2}\chi^2 - \frac{1}{2}N \log(2\pi\sigma^2) \quad (106)$$

was used in each of the 18 simulations, where σ^2 is the noise variance, N is the number of datapoints, and χ^2 is the sum of the squared residuals divided by σ^2 . The noise variance was treated as a free parameter to be estimated by MultiNest. Often, stars display short-period variability induced by oscillations, starspots, and other

Table 2

MultiNest log-evidences for 18 different models applied to the photometric signal of KOI-13b. Each model is named after the effects that it takes into account (Reflection – R, Doppler Beaming – B, Ellipsoidal Variations – E, Thermal Emissions – T). The models most favored to describe the data are in bold. Note that with the reflectance and thermal emissions models used in this study, reflected light intensity and thermal emissions cannot be distinguished in a circular orbit. For this reason, that specific case was not analyzed. The χ^2 values for the best fit eccentric models and the number of parameters for each are also listed. The two most probable models correspond to the best fit models according to the χ^2 criterion. The Null Planet model consisted of two model parameters: the noise level $\sigma \in [10^{-6}, 10^{-4}]$ and the baseline flux $\in [-0.1, 0.1]$. Last, note that the log-evidence values presented are positive due to the fact that the noise variance σ^2 is very small for this data with σ values being as low as 10^{-6} . This results in a large positive value for the second term of the log likelihood (106), which dominates the evidence integral.

Model	Circular	Eccentric	χ^2 (ppm)	Model parameters
R	37108.0 ± 0.4	37659.0 ± 5.4	2023	7
B	36970.0 ± 4.0	37166.0 ± 1.9	2539	7
E	36555.0 ± 0.5	37581.0 ± 0.4	3627	7
R + B	37108.0 ± 0.5	37670.0 ± 2.9	2018	8
R + E	37701.0 ± 0.5	37704.0 ± 2.7	2010	8
B + E	36577.0 ± 0.8	37634.0 ± 2.8	3534	7
R + B + E	37703.0 ± 1.1	37748.0 ± 1.1	1862	8
T + B + E	37703.0 ± 1.1	37764.0 ± 8.3	1817	9
R + B + E + T	...	37765.0 ± 0.9	1818	10
Null	36143.0 ± 1.0			2

effects. This variability can lead to correlated (red) noise, which can deviate from a Gaussian distribution. In that case, one may adopt a more detailed likelihood function that utilizes a nearest-neighbor approach to deal with noise correlations in the time series signal [17].

Each model was applied twice for circular and eccentric orbits. The simpler models are shown at the top of Table 2 and they increase in complexity moving down the table. The two models most favored to describe the data are those including thermal emission, Doppler boosting, and ellipsoidal variations ($\log Z = 37764 \pm 8.3$), and reflection, thermal emission, Doppler boosting, and ellipsoidal variations ($\log Z = 37765.0 \pm 0.9$), which is illustrated in Fig. 8. Based on the uncertainties on the log-evidences, these two models have an essentially equal probability to describe the observed data. This also means that adding the reflection effect to thermal emissions does not yield a significantly better fit as indicated by the χ^2 values, which indicate a difference in the sum of the squared residuals of only 0.12%. This is to be expected for planets similar to KOI-13b since they have very low albedos and are very hot due to the proximity to the host star. In each case, the eccentric model is more favored than the circular.

The astute reader will note that the log-evidence values in Table 2 are positive, which indicates that there are large positive log likelihood values in the integral. Since the likelihood, and hence the evidence, are density functions and have units, this is a result of the choice of units for flux. The log likelihood (106) is the sum of two terms. The first term is unit-less, whereas the second term depends on $\log(\sigma)$, which can change signs depending on the units.⁵ As a check, these can be estimated by considering the Null Planet model with zero baseline and a noise level of $\sigma = 5 \times 10^{-5}$. In this case, one can use the fact that there are $N = 4187$ flux data points with a standard deviation of 4.3050×10^{-5} resulting in $\chi^2 = 3104$ and a log likelihood (106) of 36066 for those particular model parameter values. Since this is the logarithm of the likelihood, large positive values like this dominate the evidence integral resulting in a log-evidence of $\log Z = 36143.0 \pm 1.0$ for the Null Planet model. Since there is a planet present, this represents

⁵ See <http://blog.stata.com/2011/02/16/positive-log-likelihood-values-happen/> for more information on this effect.

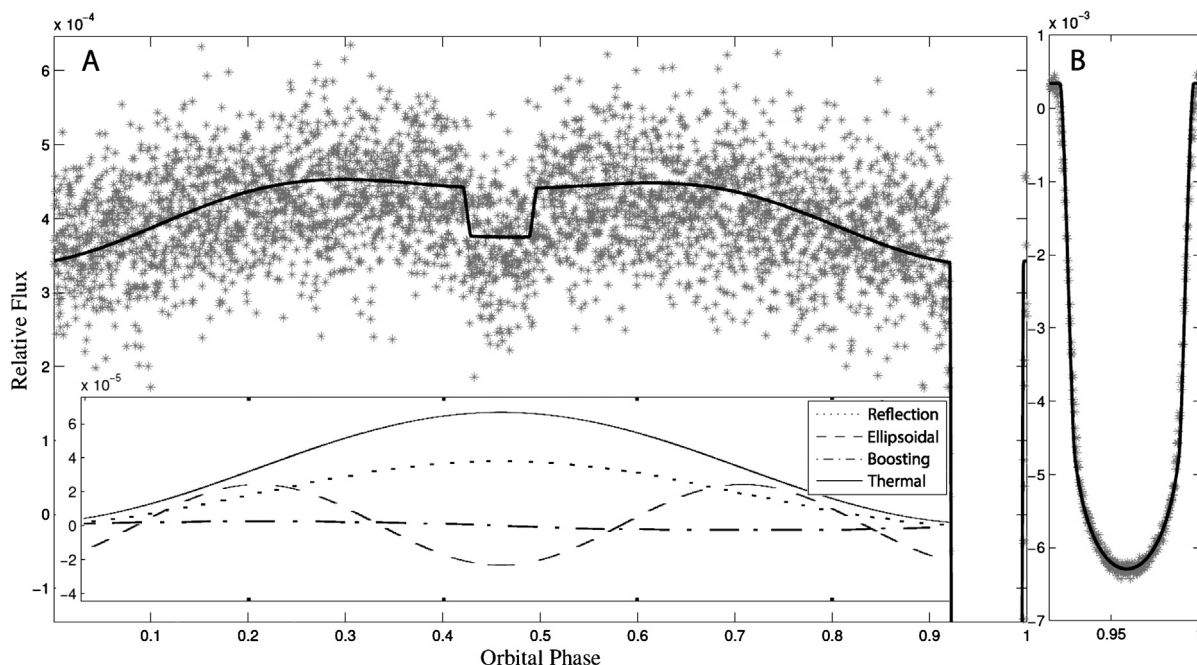


Fig. 8. A: A model fit (solid curve) to exoplanet KOI-13b data (asterisks) from the Kepler Space Telescope. The secondary eclipse from the planet passing behind the star is centered in the plot between the phases 0.4 and 0.5. The primary transit from the planet passing in front of the star occurs at the far right between phases 0.9 and 1.0. The inset at the bottom shows the estimated photometric flux contributions from reflected light, ellipsoidal variations, Doppler boosting and thermal emissions. B: A detailed illustration of the model fit to the primary transit.

a lower bound to the positive log-evidence values obtained in Table 2.

By comparing the results from multiple models some important facts about the KOI-13b are revealed. First, both the χ^2 values and the log-evidences indicate that reflected light (or thermal emissions, which are similar to reflected light) is a prominent component in the photometric signature. One can also note that the R + E model, which describes reflection and ellipsoidal variations ($\log Z = 37704 \pm 2.7$), is only marginally less probable than the models that include all three photometric effects. This further implies that ellipsoidal variations also play a significant role in the observed data. This indicates an additional advantage to comparing and contrasting sets of models based on the χ^2 values and the log-evidences. By calculating the Bayesian evidence and incorporating model testing by turning on and off certain photometric effects, one can effectively characterize planetary systems, as well as use it as a planetary confirmation procedure.

5.4. Force field selection in biomolecular structure determination

Last, we present an example from Habeck in structural biology [124]. NMR spectroscopy allows us to determine the three-dimensional structure of complex biomolecules such as proteins at atomic resolution. However, often the data are not sufficient to determine the structure without additional guidance from molecular mechanics force fields. These force fields can be very complex, which slows down the structure calculation. Therefore, the force fields used in biomolecular structure determination typically neglect important contributions such as electrostatic or solvent interactions and rather work with a minimalist force field. On the other hand it is clear that by choosing more realistic force fields the results obtained from challenging data will be more useful.

Current practice to calculate biomolecular structures is to set up a cost function (the so-called hybrid energy) $\lambda D(x, d) + E(x)$ that is comprised of a data fitting term $D(x, d)$ weighted by λ and a force field $E(x)$ where x are the conformational degrees of freedom of the biomolecule (e.g. the Cartesian coordinates of all atoms or

dihedral angles) and d represents relevant data. Inferential structure determination (ISD) [125] is a strictly probabilistic approach to solve structure determination problems. It not only allows us to estimate the appropriate weight of the data λ [126], but also to compare two alternative force fields in the light of given experimental data [124] as well as determine the best weight of the force field [127]. ISD models the data d probabilistically such that

$$P(d|x, M, I) = \frac{1}{Z_D(\lambda, d)} e^{-\lambda D(x, d)} \quad (107)$$

where $Z_D(\lambda, d)$ is a normalizing constant that depends on the chosen model M to assess discrepancies between observed data d and predictions made by the forward model. The force field $E(x)$ is incorporated using a Boltzmann distribution as prior probability over the conformational degrees of freedom:

$$P(x|M, I) = \frac{1}{Z_E(\beta)} e^{-\beta E(x)} \quad (108)$$

where $Z_E(\beta)$ is the partition function of the Boltzmann distribution and normalizes the prior. In the most general case the inverse temperature β of the force field is unknown because, as explained above, we cannot afford to work with realistic force fields but have to make drastic simplifications. Therefore also the “temperature” of the minimalist force field is no longer identical to the temperature at which the experiments were carried out, but is instead an unknown hyperparameter [127].

Here, we compare two different force fields that are used in biomolecular modeling. Both aim to describe van der Waals interactions between atoms that are not linked via a covalent bond. The first is a quartic repulsion term that drops to zero when the distance between two atoms r_{ij} is larger than the sum of their van der Waals radii R_i [128]:

$$E_{\text{quartic}}(r_{ij}) = \begin{cases} (r_{ij} - R_i - R_j)^4, & r_{ij} \leq R_i + R_j \\ 0; & r_{ij} > R_i + R_j \end{cases} \quad (109)$$

This force field ignores the attractive contribution of the van der Waals interaction. An alternative force field that takes the attrac-

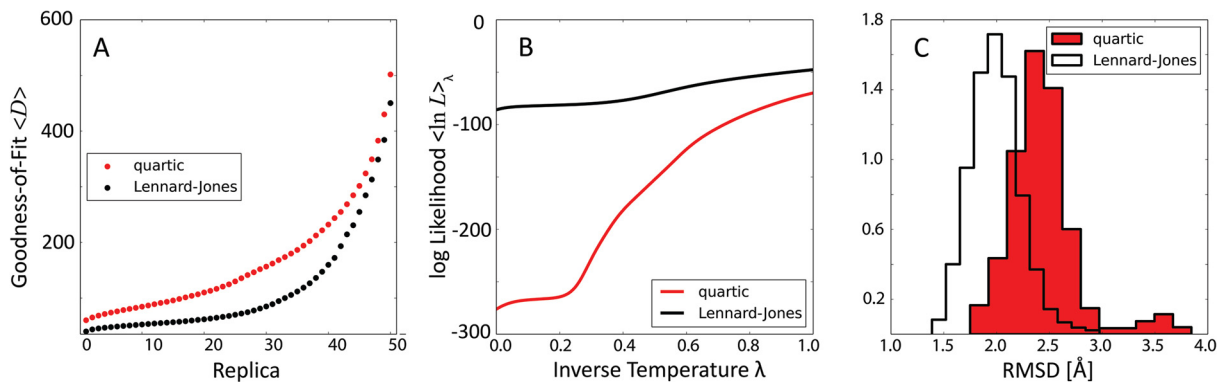


Fig. 9. Comparison of force fields in biomolecular structure determination. A: Average goodness of fit $\langle D \rangle_{\lambda,q}$ in the 50 replicas at varying λ and q used to sample the posterior. B: Average log likelihood $\langle \log L \rangle_{\lambda}$ obtained with the estimated DOS. C: Accuracy as measured by the root mean square deviation (RMSD) of the structures sampled from the posterior with the crystal structure.

tive term into account is used in the Rosetta software [129]. This is a Lennard–Jones potential that is linearly ramped to finite values as r_{ij} approaches zero and vanishes for distances larger than a cutoff distance of $R_{\text{cut}} = 5.5$ Å. The potential function is:

$$E_{\text{LJ}}(r_{ij}) = \left(\frac{R_i + R_j}{r_{ij}} \right)^{12} - 2 \left(\frac{R_i + R_j}{r_{ij}} \right)^6 \quad (110)$$

for $0.6(R_i + R_j) < r_{ij} \leq 5$ Å and continues linearly to the left and right of this interval.

We compare these two force fields in the light of NMR data measured on the Fyn-SH3 domain, a small signaling domain of 59 amino acids in length. The data are sparse and comprised of 154 inter-proton distances measured on a deuterated sample [125]. We ran a parallel tempering simulation for each of the two force fields. The parallel tempering schedule is two-dimensional [130]: The first replica parameter λ is the inverse temperature and gradually switches off the data, whereas the second parameter is Tsallis' q used to deviate from the Boltzmann ensemble (108). The Tsallis ensemble approaches the Boltzmann ensemble for $q \rightarrow 1$ and is used here only for convenience because neighboring replicas will show a higher overlap due to the fatter tails of the Tsallis ensemble. Fifty replicas were set up in which λ varied from 0.1 to 1.0 and q varied from 1.06 to 1.0; we used the same combination of (λ, q) values for both force fields.

Fig. 9A shows the average goodness of fit $\langle D \rangle_{\lambda,q}$ (negative log likelihood) for each of the 50 replicas. It is already apparent from this figure that the Lennard–Jones potential (110) results in a better goodness of fit than the purely repulsive potential [Eq. (109)]. We applied histogram re-weighting to estimate the density of states (DOS) from the replica simulations [82,89]. The estimated DOS can be used to calculate the expected log-likelihood as a function of the inverse temperature $\langle \log L \rangle_{\lambda}$ and apply thermodynamic integration, which would not be possible without the help of the DOS because (λ, q) are varied simultaneously. Fig. 9B shows the expected log likelihood $\langle \log L \rangle_{\lambda}$ as a function of the inverse temperature, i.e. the integrand of thermodynamic integration equation (49). Alternatively, we can evaluate the partition function (41) to compute the evidence. Both approaches are equivalent and give the same result. The evidence clearly favors the Lennard–Jones potential ($\log Z = -69$) over the potential based on a quartic repulsion term ($\log Z = -166$). The Lennard–Jones potential is not only more supported by the NMR data but also results in a more accurate structure ensemble. The root mean square deviation (RMSD) between members of the posterior ensemble and the crystal structure, serving here as a reference, is systematically shifted towards better values when using the Lennard–Jones potential (see Fig. 9C).

6. Conclusion

In this paper we have reviewed the concept of the Bayesian evidence (marginal likelihood) and the related concepts of Bayes factors and odds ratios, which quantify the probability of one model over another based on the selected models and the data. That is, the degree to which the data implies a given model. In addition to discussing the analytic treatment of the foundations, we have focused mainly on approximate and numerical techniques such as the Laplace approximation, variational Bayes, thermodynamic integration and stochastic integration via Monte Carlo methods.

The discussions regarding these methods were supplemented by four examples with the intention of demonstrating Bayesian model testing in different scientific domains: signal detection (BCI) [115], sensor characterization (robotics) [121], scientific model selection (exoplanet characterization) [122,123] and molecular force characterization (structural biology) [124]. Together these applications demonstrate the power of Bayesian model testing in a variety of contexts leading to improved signal processing algorithms, improved instrument models, as well as a deeper understanding of physical systems at scales ranging from the astronomic to the microscopic. These examples, which involve detailed theoretical signal models, do not begin to cover the vast array of inference problems and underlying models that one could consider. For example, nonparametric models find great use in domains where detailed signal models are lacking. Examples of such models include Gaussian Processes [131,132,76,133] and generalized autoregressive models [100,74]. As demonstrated in the provided references, Bayesian model testing works well with those nonparametric models as well.

In the examples given in this paper, model selection was based on the evidence or Bayes factors alone. However, it is important to remember that probability theory is not decision theory [57]. That is, there are other factors involved in any decision-making process that can be described by a utility function that maximizes expected utility or minimizes expected loss. For this reason, it is strongly recommended that model selection be performed by considering both the probability of a model and the expected utility function [134]. In practice, this can quite challenging as it is often difficult to identify and to quantify such utility especially in situations where there are multiple factors involved. This remains an active area of research.

Acknowledgments

This paper was written for a special issue dedicated to William J. Fitzgerald. While only once did I (Knuth) have the pleasure to meet Prof. Fitzgerald, I was influenced by him indirectly through

several of his students, Ali Taylan Cemgil, Ercan E. Kuruoglu, and Robin D. Morris, who in addition to always speaking very highly of him were each clearly excellently trained as creative, talented and responsible members of the scientific community. This research was supported in part by a University at Albany Faculty Research Awards Program (FRAP-A) Award (Knuth), a Deutsche Forschungsgemeinschaft (DFG) grant HA 5918/1-1 (Habeck) as well as two University at Albany Benevolent Research Grant Awards (Malakar and Mubeen). Additional N. Malakar support carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. The authors would like to thank the anonymous reviewers for their careful, thorough, and extensive comments, which have served to improve the quality of this manuscript.

References

- [1] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, 2003.
- [2] P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Vol. 10, Cambridge University Press, Cambridge, UK, 2005.
- [3] J.V. Candy, *Bayesian Signal Processing: Classical, Modern and Particle Filtering Methods*, John Wiley & Sons, Inc., 2009.
- [4] U. von Toussaint, Bayesian inference in physics, *Rev. Mod. Phys.* 83 (3) (2011) 943.
- [5] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, *Bayesian Data Analysis*, 3rd edition, CRC Press, 2013.
- [6] W. von der Linden, V. Dose, U. von Toussaint, *Bayesian Probability Theory: Applications in the Physical Sciences*, Cambridge University Press, 2014.
- [7] H. Jeffreys, *Theory of Probability*, 3rd edition, Oxford University Press, Oxford, 1961.
- [8] M.H. DeGroot, *Optimal Statistical Decisions*, Vol. 82, John Wiley & Sons, 2004.
- [9] R.E. Kass, A.E. Raftery, Bayes factors, *J. Am. Stat. Assoc.* 90 (430) (1995) 773–795.
- [10] A. Zellner, A. Siow, Posterior odds ratios for selected regression hypotheses, *Trab. Estad. Investig. Oper.* 31 (1) (1980) 585–603.
- [11] D.S. Sivia, W.I.F. David, K.S. Knight, S.F. Gull, An introduction to Bayesian model selection, *Physica D* 66 (1) (1993) 234–242.
- [12] C. Andrieu, A. Doucet, W.J. Fitzgerald, J.M. Pérez, Bayesian computational approaches to model selection, in: W.J. Fitzgerald, R.L. Smith, A.T. Walden, P.C. Young (Eds.), *Nonlinear and Nonstationary Signal Processing*, Cambridge, 1998, Cambridge University Press, Cambridge, 1998, pp. 1–41.
- [13] W.J. Fitzgerald, Markov chain Monte Carlo methods with applications to signal processing, *Signal Process.* 81 (1) (2001) 3–18.
- [14] J.M. Bernardo, A.F.M. Smith, *Bayesian Theory*, Vol. 405, John Wiley & Sons, 2009.
- [15] D.J.C. MacKay, Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks, *Netw. Comput. Neural Syst.* 6 (3) (1995) 469–505.
- [16] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, 2003.
- [17] D.S. Sivia, *J. Skilling, Data Analysis: A Bayesian Tutorial*, 2nd edition, Oxford University Press, Oxford, 2006.
- [18] N. Xiang, P.M. Goggans, Evaluation of decay times in coupled spaces: Bayesian decay model selection, *J. Acoust. Soc. Am.* 113 (5) (2003) 2685–2697.
- [19] T. Jasa, N. Xiang, Nested sampling applied in Bayesian room-acoustics decay analysis, *J. Acoust. Soc. Am.* 132 (5) (2012) 3251–3262.
- [20] P. Goggans, R.W. Henderson, N. Xiang, Using nested sampling with Galilean Monte Carlo for model comparison problems in acoustics, *Proc. Meet. Acoust.* 19 (1) (2013), <http://dx.doi.org/10.1121/1.4800876>, <http://scitation.aip.org/content/asa/journal/poma/19/1/10.1121/1.4800876>.
- [21] T.J. Loredo, D.Q. Lamb, Bayesian analysis of neutrinos observed from supernova SN 1987A, *Phys. Rev. D* 65 (6) (2002) 063002.
- [22] A.R. Liddle, P. Mukherjee, D. Parkinson, Y. Wang, Present and future evidence for evolving dark energy, *Phys. Rev. D* 74 (12) (2006) 123506.
- [23] J. Clark, I.S. Heng, M. Pitkin, G. Woan, Evidence-based search method for gravitational waves from neutron star ring-downs, *Phys. Rev. D* 76 (4) (2007) 043003.
- [24] A.F. Heavens, T.D. Kitching, L. Verde, On model selection forecasting, dark energy and modified gravity, *Mon. Not. R. Astron. Soc.* 380 (3) (2007) 1029–1035.
- [25] A.R. Liddle, Information criteria for astrophysical model selection, *Mon. Not. R. Astron. Soc., Lett.* 377 (1) (2007) L74–L78.
- [26] E. Jullo, J.-P. Kneib, M. Limousin, Á. Elíasdóttir, P.J. Marshall, T. Verdugo, A Bayesian approach to strong lensing modelling of galaxy clusters, *New J. Phys.* 9 (12) (2007) 447.
- [27] R. Trotta, Bayes in the sky: Bayesian inference and model selection in cosmology, *Contemp. Phys.* 49 (2) (2008) 71–104.
- [28] A.R. Liddle, Statistical methods for cosmological parameter selection and estimation, arXiv preprint arXiv:0903.4210.
- [29] F. Feroz, S.T. Balan, M.P. Hobson, Bayesian evidence for two companions orbiting HIP 5158, *Mon. Not. R. Astron. Soc., Lett.* 416 (1) (2011) L104–L108.
- [30] F. Feroz, Calculation and applications of Bayesian evidence in astrophysics and particle physics phenomenology, in: 2013 IEEE 13th International Conference on Data Mining Workshops, ICDMW, IEEE, 2013, pp. 8–15.
- [31] I. Debono, Bayesian model selection for dark energy using weak lensing forecasts, *Mon. Not. R. Astron. Soc.* 437 (1) (2014) 887–897.
- [32] D.S. Sivia, C.J. Carlile, Molecular spectroscopy and Bayesian spectral analysis—how many lines are there?, *J. Chem. Phys.* 96 (1) (1992) 170–178.
- [33] M.J. Beal, Z. Ghahramani, The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures, in: J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, M. West (Eds.), *Bayesian Statistics 7*, in: *Bayesian Statistics*, vol. 7, Oxford University Press, Oxford, 2003, pp. 453–464.
- [34] P. Marshall, N. Rajguru, A. Slosar, Bayesian evidence as a tool for comparing datasets, *Phys. Rev. D* 73 (6) (2006) 067302.
- [35] D. MacKay, The evidence framework applied to classification networks, *Neural Comput.* 4 (5) (1992) 720–736.
- [36] D.H. Wolpert, On the use of evidence in neural networks, *Adv. Neural Inf. Process. Syst.* (1993) 539–546.
- [37] W.D. Penny, S.J. Roberts, Bayesian neural networks for classification: how useful is the evidence framework?, *Neural Netw.* 12 (6) (1999) 877–892.
- [38] M.S. Lewicki, Bayesian modeling and classification of neural signals, *Neural Comput.* 6 (5) (1994) 1005–1030.
- [39] N.J. Trujillo-Barreto, E. Aubert-Vázquez, P.A. Valdés-Sosa, Bayesian model averaging in EEG/MEG imaging, *NeuroImage* 21 (4) (2004) 1300–1319.
- [40] M.W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, S.M. Smith, Bayesian analysis of neuroimaging data in FSL, *Neuroimage* 45 (1) (2009) S173–S186.
- [41] K. Friston, W. Penny, Post hoc Bayesian model selection, *Neuroimage* 56 (4) (2011) 2089–2099.
- [42] S. Gulam Razul, W.J. Fitzgerald, C. Andrieu, Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC, *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* 497 (2) (2003) 492–510.
- [43] L. De Cruz, D.G. Ireland, P. Vancraeyveld, J. Ryckebusch, Bayesian model selection for electromagnetic kaon production on the nucleon, *Int. J. Mod. Phys. A* 26 (03n04) (2011) 642–644.
- [44] J. Bergström, Bayesian evidence for non-zero θ_{13} and CP-violation in neutrino oscillations, *J. High Energy Phys.* 2012 (8) (2012) 1–20.
- [45] A. Nallanathan, W.J. Fitzgerald, Bayesian model selection applied to spatial signal processing, *IEE Proc., Vis. Image Signal Process.* 141 (1) (1994) 76–80.
- [46] S.J. Roberts, Independent component analysis: source assessment and separation, a Bayesian approach, *IEE Proc., Vis. Image Signal Process.* 145 (3) (1998) 149–154.
- [47] B. Kannan, W.J. Fitzgerald, E.E. Kuruoglu, Joint DOA, frequency and model order estimation in additive α -stable noise, in: *ICASSP'00 Proc.* 6, Vol. 6, IEEE, 2000, pp. 3798–3801.
- [48] S. Roberts, R. Everson, *Independent Component Analysis: Principles and Practice*, Cambridge University Press, 2001.
- [49] C. Andrieu, P.M. Djurić, A. Doucet, Model selection by MCMC computation, *Signal Process.* 81 (1) (2001) 19–37.
- [50] W.D. Penny, S.J. Roberts, Bayesian multivariate autoregressive models with structured priors, *IEE Proc., Vis. Image Signal Process.* 149 (1) (2002) 33–41.
- [51] E. Puskaya, C. Andrieu, A. Doucet, W.J. Fitzgerald, Bayesian curve fitting using MCMC with applications to signal segmentation, *IEEE Trans. Signal Process.* 50 (3) (2002) 747–758.
- [52] J.L. Beck, K.-V. Yuen, Model selection using response measurements: Bayesian probabilistic approach, *J. Eng. Mech.* 130 (2) (2004) 192–203.
- [53] C. Simon, P. Weber, E. Levrat, Bayesian networks and evidence theory to model complex systems reliability, *J. Comput.* 2 (1) (2007) 33–43.
- [54] S. Chulani, B. Boehm, B. Steece, Bayesian analysis of empirical software engineering cost models, *IEEE Trans. Softw. Eng.* 25 (4) (1999) 573–583.
- [55] D. Madigan, A.E. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window, *J. Am. Stat. Assoc.* 89 (428) (1994) 1535–1546.
- [56] J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, *Bayesian model averaging: a tutorial*, *Stat. Sci.* (1999) 382–401.
- [57] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer Verlag, New York, 1985.
- [58] G.E.P. Box, G.C. Tiao, *Bayesian Inference in Statistical Analysis*, John Wiley & Sons, New York, 1992.
- [59] C. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd edition, Springer Verlag, New York, 2007.
- [60] R.T. Cox, Probability, frequency, and reasonable expectation, *Am. J. Phys.* 14 (1946) 1–13.

- [61] R.T. Cox, *The Algebra of Probable Inference*, Johns Hopkins Press, Baltimore, 1961.
- [62] K.H. Knuth, Measuring on lattices, in: P. Goggans, C.-Y. Chan (Eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Oxford, MS, USA, in: AIP Conf. Proc., vol. 1193, AIP, New York, 2009, pp. 132–144, arXiv:0909.3684v1 [math.GM], 2009.
- [63] K.H. Knuth, J. Skilling, *Foundations of inference*, *Axioms* 1 (1) (2012) 38–73, <http://www.mdpi.com/2075-1680/1/1/38>.
- [64] G.L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Vol. 48, Springer, Berlin, 1988.
- [65] S. Chib, Marginal likelihood from the Gibbs output, *J. Am. Stat. Assoc.* 90 (432) (1995) 1313–1321.
- [66] J. Berkhof, I. Van Mechelen, A. Gelman, A Bayesian approach to the selection and testing of mixture models, *Stat. Sin.* 13 (2) (2003) 423–442.
- [67] M. Trias, A. Vecchio, J. Veitch, Delayed rejection schemes for efficient Markov-chain Monte-Carlo sampling of multimodal distributions, arXiv:0904.2207 [stat.ME], 2009.
- [68] S. Chib, S. Ramamurthy, Tailored randomized block MCMC methods with application to DSGE models, *J. Econom.* 155 (1) (2010) 19–38.
- [69] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* 82 (4) (1995) 711–732.
- [70] S.P. Brooks, P. Giudici, G.O. Roberts, Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions, *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 65 (1) (2003) 3–39.
- [71] H.F. Lopes, M. West, Bayesian model assessment in factor analysis, *Stat. Sin.* 14 (1) (2004) 41–68.
- [72] C. Han, B.P. Carlin, Markov chain Monte Carlo methods for computing Bayes factors, *J. Am. Stat. Assoc.* 96 (2001) 455.
- [73] H.E. Daniels, Saddlepoint approximations in statistics, *Ann. Math. Stat.* (1954) 631–650.
- [74] W. Penny, S. Kiebel, K. Friston, *Variational Bayes*, in: K. Friston, J. Ashburner, S. Kiebel, T. Nichols, W. Penny (Eds.), *Statistical Parametric Mapping: the Analysis of Functional Brain Images*, Elsevier, London, 2006, pp. 303–312.
- [75] S.F. Gull, Developments in maximum entropy data analysis, in: J. Skilling (Ed.), *Maximum Entropy and Bayesian Methods*, Vol. 36, Springer, 1989, pp. 53–71.
- [76] C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.
- [77] H. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *J. R. Stat. Soc. B* 71 (2) (2009) 319–392.
- [78] T.G. Martins, D. Simpson, F. Lindgren, H. Rue, Bayesian computing with INLA: new features, *Comput. Stat. Data Anal.* 67 (2013) 68–83.
- [79] T.P. Minka, Expectation propagation for approximate Bayesian inference, in: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.
- [80] B. Cseke, T. Heskes, Approximate marginals in latent Gaussian models, *J. Mach. Learn. Res.* 12 (2011) 417–454.
- [81] R.M. Neal, Probabilistic inference using Markov chain Monte Carlo methods, tech. rep. CRG-TR-93-1, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 1993.
- [82] M. Habeck, Evaluation of marginal likelihoods using the density of states, in: N. Lawrence, M. Girolami (Eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, *J. Mach. Learn. Res. Workshop Conf. Proc.* 22 (2012) 486–494.
- [83] A. Gelman, X. Meng, Simulating normalizing constants: from importance sampling to bridge sampling to path sampling, *Stat. Sci.* 13 (1998) 163–185.
- [84] R.H. Swendsen, J.-S. Wang, Replica Monte Carlo simulation of spin glasses, *Phys. Rev. Lett.* 57 (1986) 2607–2609.
- [85] C.J. Geyer, Markov chain Monte Carlo maximum likelihood, in: E.M. Keramidas, S.M. Kaufman (Eds.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 1991, pp. 156–163.
- [86] R.M. Neal, Annealed importance sampling, *Stat. Comput.* 11 (2001) 125–139.
- [87] J. Skilling, Nested sampling, in: R. Fischer, V. Dose, R. Preuss, U. von Toussaint (Eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Garching, Germany 2004, in: AIP Conf. Proc., vol. 735, AIP, New York, 2004, pp. 395–405, <http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.1835238>, 2004.
- [88] J. Skilling, Nested sampling for general Bayesian computation, *Bayesian Anal.* 1 (4) (2006) 833–859.
- [89] M. Habeck, Bayesian estimation of free energies from equilibrium simulations, *Phys. Rev. Lett.* 109 (10) (2012) 100601.
- [90] M. Habeck, Ensemble annealing of complex systems, arXiv:1504.00053 [physics.comp-ph], 2015.
- [91] S. Kirkpatrick, C.D.G. Gelatt, M.P. Vecchi, Optimization by simulated annealing, *Science* 220 (4598) (1983) 671–680.
- [92] G.E. Crooks, Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences, *Phys. Rev. E* 60 (1999) 2721–2726.
- [93] G. Hinton, D. Van Camp, Keeping the neural networks simple by minimizing the description length of the weights, in: *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, ACM, 1993, pp. 5–13.
- [94] D.J.C. MacKay, Developments in probabilistic modelling with neural networks-ensemble learning, in: B. Kappen, S. Gielen (Eds.), *Neural Networks: Artificial Intelligence and Industrial Applications*, Springer, 1995, pp. 191–198.
- [95] H. Valpola, J. Karhunen, An unsupervised ensemble learning method for nonlinear dynamic state-space models, *Neural Comput.* 14 (11) (2002) 2647–2692.
- [96] S. Waterhouse, D. MacKay, T. Robinson, Bayesian methods for mixtures of experts, *Adv. Neural Inf. Process. Syst.* (1996) 351–357.
- [97] H. Attias, Inferring parameters and structure of latent variable models by variational Bayes, in: *Proc. of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., 1999, pp. 21–30.
- [98] N. Lawrence, C. Bishop, *Variational Bayesian independent component analysis*, tech. report, Univ. of Cambridge, 2000.
- [99] M.-A. Sato, Online model selection based on the variational Bayes, *Neural Comput.* 13 (7) (2001) 1649–1681.
- [100] S. Roberts, W. Penny, Variational Bayes for generalized autoregressive models, *IEEE Trans. Signal Process.* 50 (9) (2002) 2245–2257.
- [101] V. Šmídl, A. Quinn, *The Variational Bayes Method in Signal Processing*, Springer Science & Business Media, 2006.
- [102] R.P. Feynman, *Statistical Mechanics: A Set of Lectures*, Frontiers in Physics, W.A. Benjamin, 1972.
- [103] W. McComb, *Renormalization Methods: A Guide for Beginners*, Clarendon Press, 2004.
- [104] Z. Ghahramani, M.J. Beal, Propagation algorithms for variational Bayesian learning, *Adv. Neural Inf. Process. Syst.* (2001) 507–513.
- [105] J.M. Winn, C.M. Bishop, Variational message passing, *J. Mach. Learn. Res.* 6 (2005) 661–694.
- [106] T.S. Jaakkola, M.I. Jordan, Computing upper and lower bounds on likelihoods in intractable networks, in: *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., 1996, pp. 340–348.
- [107] P. Mukherjee, D. Parkinson, A.R. Liddle, A nested sampling algorithm for cosmological model selection, *Astrophys. J. Lett.* 638 (2) (2006) L51.
- [108] F. Feroz, M.P. Hobson, Multimodal nested sampling: an efficient and robust alternative to Markov chain Monte Carlo methods for astronomical data analyses, *Mon. Not. R. Astron. Soc.* 384 (2) (2008) 449–463.
- [109] F. Feroz, M.P. Hobson, M. Bridges, MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics, *Mon. Not. R. Astron. Soc.* 398 (4) (2009) 1601–1614.
- [110] M. Betancourt, Nested sampling with constrained Hamiltonian Monte Carlo, arXiv:1005.0157 [physics.data-an], 2010.
- [111] J. Skilling, Bayesian computation in big spaces-nested sampling and Galilean Monte Carlo, in: P. Goyal, A. Giffin, K.H. Knuth, E. Vrscay (Eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Jackson Hole WY, USA 2003, vol. 1443, AIP, New York, 2003, pp. 145–156, <http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.3703630>.
- [112] F. Feroz, J. Skilling, Exploring multi-modal distributions with nested sampling, in: U. von Toussaint, R. Fischer (Eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Garching, Germany, 2012, in: AIP Conf. Proc., vol. 1553, AIP, New York, 2013, pp. 106–113, arXiv:1312.5638 [astro-ph.IM].
- [113] M. Habeck, Nested sampling with demons, in: A. Mohammad-Djafari, F. Barbaresco (Eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Amboise, France, 2014, in: AIP Conf. Proc., vol. 1641, AIP, New York, 2014, pp. 121–129.
- [114] B.J. Brewer, L.B. Pártay, G. Csányi, Diffusive nested sampling, *Stat. Comput.* 21 (4) (2011) 649–656.
- [115] A.M. Mubeen, K.H. Knuth, Bayesian odds-ratio filters: a template-based method for online detection of P300 evoked responses, arXiv:1304.1565 [q-bio.NC], 2014.
- [116] M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, 1972.
- [117] E.W. Sellers, D.J. Krusienski, D.J. McFarland, T.M. Vaughan, J.R. Wolpaw, A P300 event-related potential brain-computer interface (BCI): the effects of matrix size and inter stimulus interval on performance, *Biol. Psychol.* 73 (3) (2006) 242–252.
- [118] N. Yeung, R. Bogacz, C.B. Holroyd, S. Nieuwenhuis, J.D. Cohen, Generation of simulated EEG data, <http://www.cs.bris.ac.uk/~rafal/phasereset/>, 2006 [Online; accessed 19-Oct-2014].
- [119] K.H. Knuth, H.G.J. Vaughan, The Bayesian origin of blind source separation and electromagnetic source estimation, in: W. von der Linden, V. Dose, R. Fischer, R. Preuss (Eds.), *Maximum Entropy and Bayesian Methods*, Munich 1998, Kluwer, Dordrecht, 1999, pp. 217–226.
- [120] N.A. Obuchowski, ROC analysis, *Am. J. Roentgenol.* 184 (2) (2005) 364–372.
- [121] N.K. Malakar, D. Gladkov, K.H. Knuth, Modeling a sensor to improve its efficiency, *J. Sens.* 2013 (2013) 11, <http://dx.doi.org/10.1155/2013/481054>.
- [122] K.H. Knuth, B. Placek, Z. Richards, Detection and characterization of non-transiting extra-solar planets in Kepler data using reflected light variations, in: N. Chawla, A.N. Srivastava (Eds.), *Proceedings of the Conference on Intelligent Data Understanding 2012 (IEEE Explore)*, 2012, pp. 31–38.

- [123] B. Placek, K.H. Knuth, D. Angerhausen, EXONEST: Bayesian model selection applied to the detection and characterization of exoplanets via photometric variations, *Astrophys. J.* 795 (2) (2014) 112, arXiv:1310.6764 [astro-ph.EP], <http://stacks.iop.org/0004-637X/795/i=2/a=112>.
- [124] M. Habeck, Statistical mechanics analysis of sparse data, *J. Struct. Biol.* 173 (2011) 541–548.
- [125] W. Rieping, M. Habeck, M. Nilges, Inferential Structure Determination, *Science* 309 (2005) 303–306.
- [126] M. Habeck, W. Rieping, M. Nilges, Weighting of experimental evidence in macromolecular structure determination, *Proc. Natl. Acad. Sci. USA* 103 (2006) 1756–1761.
- [127] M. Mechelke, M. Habeck, Calibration of Boltzmann distribution priors in Bayesian data analysis, *Phys. Rev. E* 86 (2012) 066705.
- [128] M. Habeck, M. Nilges, W. Rieping, Bayesian inference applied to macromolecular structure determination, *Phys. Rev. E* 72 (2005) 031912.
- [129] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Design of a novel globular protein fold with atomic-level accuracy, *Science* 302 (2003) 1364–1368.
- [130] M. Habeck, M. Nilges, W. Rieping, Replica-exchange Monte Carlo scheme for Bayesian data analysis, *Phys. Rev. Lett.* 94 (2005) 0181051.
- [131] M.N. Gibbs, Bayesian Gaussian processes for regression and classification, Ph.D. thesis, University of Cambridge, 1998.
- [132] D.J.C. MacKay, Introduction, in: *Gaussian Processes*, in: NATO ASI Series F Computer and Systems Sciences, vol. 168, 1998, pp. 133–166.
- [133] M. Osborne, Bayesian Gaussian processes for sequential prediction, optimisation and quadrature, Ph.D. thesis, Oxford University New College, 2010.
- [134] D.W. Hogg, Data analysis recipes: making decisions, <https://github.com/davidwhogg/DataAnalysisRecipes/blob/master/decision/decision.tex>, 2013.

Kevin H. Knuth is an Associate Professor in the Departments of Physics and Informatics at the University at Albany, Albany NY USA. He is Editor-in-Chief of the journal *Entropy*, and is the co-founder and President of a robotics company, Autonomous Exploration Inc, and a former NASA research scientist. He has 20 years of experience in applying Bayesian and maximum entropy methods to the design of machine learning algorithms for data analysis applied to astronomy and the physical sciences. His current research interests include the foundations of physics, autonomous robotics, and searching for and characterizing extrasolar planets.

Michael Habeck received his Ph.D. in Biophysics in 2004 from the University of Regensburg, Germany. In 2009, he started an independent research group at the Max Planck Institute for Developmental Biology in Tübingen, Germany. Since 2013 he is a group leader in Computational Structural Biology at the University of Göttingen and the Max Planck Institute for Biophysical Chemistry. His major research interests are in the application of Bayesian inference to data analysis problems arising in structural biology.

Nabin K. Malakar received his Ph.D. degree in physics from the University at Albany (SUNY), Albany NY USA, in 2011. He is currently a postdoctoral scholar at the Jet Propulsion Laboratory, California Institute of Technology, in Pasadena. He is working on the development, validation and evaluation of new land surface temperature and emissivity products from the MODIS instrument onboard the Terra and Aqua satellites. His research interests include identification of relevant variables in a physical phenomena to improve our understanding of atmospheric processes, as well as algorithm development for various applications of remote sensing data.

Asim M. Mubeen received his M.Sc. degree in physics from the Punjab University Lahore, Pakistan in 1998, and M.S. degree in physics from the University at Albany (SUNY), Albany NY USA, in 2007. He is currently a Ph.D student in Department of Physics at University at Albany (SUNY), Albany USA. Since December 2013, he is an Assistant Research Scientist, in the Geriatrics Division at Nathan Kline Institute, Orangeburg, NY. His research interests include Bayesian inference, digital signal processing, brain computer interface, image processing, and diffusion tensor imaging (DTI).

Ben Placek received his Ph.D. in physics from the University at Albany (SUNY), Albany NY USA, in 2014 and is currently a Physics Instructor at Schenectady County Community College. His research is focused on exoplanet detection and characterization, and he has worked to develop the EXONEST algorithm, which employs Bayesian methods to improve exoplanet characterization.