Estimating probability density functions using a combined maximum entropy moments and Bayesian method. Theory and numerical examples

N. Armstrong[a], G. J. Sutton[b,c] and D. B. Hibbert[b]

[a]Defence Science and Technology
Fishermans Bend, Vic 3207, AUSTRALIA

[b]School of Chemistry,
University of New South Wales,
Sydney NSW 2052, AUSTRALIA

[c]Global Big Data Technologies Centre
University of Technology Sydney
Ultimo NSW 2007, AUSTRALIA

December 3, 2018

Abstract

Estimating the probability density function (pdf) from a limited sample of data is a challenging data analysis problem. Furthermore, determining which pdf best describes the available data involves an extra layer of complexity to the analysis, which if ignored, can have considerable consequences.

We propose a combined maximum entropy (MaxEnt) moments and Bayesian model selection method to address this problem. The MaxEnt moments component is used to formulate a set of possible pdf models, each constrained by a different set of moments and parameterised by a set of Lagrangian multipliers. The Bayesian model selection component makes an inference about the most probable model, from the set of MaxEnt moment models. The structure of the prior pdf for the Lagrangian multipliers is determined from an expansion of the free energy functional for each MaxEnt model, and corresponding hyperparamaters are calculated empirically.

Numerical experiments were used to test the proposed method on samples taken from Gaussian and (more complex) non-Gaussian distributions, over a range of sample sizes. The results clearly demonstrate that the method can discriminate between simple and complex MaxEnt models for sample sizes approximately greater than 60. Our results demonstrate that MaxEnt and Bayesian methods are complementary. More critically, Bayesian inference is necessary when a set of competing MaxEnt models can be derived for a single dataset from a range of assumptions.

Keywords: Bayesian model selection, maximum entropy, probability density function, moments

1 Introduction and Background

1.1 Introduction

There are many situations in which a probability density function (pdf) needs to be determined from a limited random sample of data. Typically the underlying physical, chemical, biological and even socio-economic mechanisms are so complex that it is difficult to derive an accurate model. Using the available data, an inference is made regarding the most plausible distribution given a set of distribution functions. Each distribution comes with specific assumptions about its shape and tails, which can impact subsequent analysis and decision making. There are also a wide range statistical methods for determining a pdf, conditional on the data and available prior information, including standard likelihood method [1, 2], Bayesian methods [3–6] and non-parametric methods, such as histogram [7–9] and kernel density estimation [10–12]. Regardless of the statistical method employed, this type of analysis is essentially a problem of missing information, where limited and noisy data is used to make an inference about the underlying pdf.

We present a combined maximum entropy (MaxEnt) method of moments with a Bayesian model selection procedure to determine the most plausible pdf conditional on the available data. The MaxEnt moments method, proposed by Mead and Papanicolaou [13], ensures the candidate pdfs belong to a wide class of distributions, without examining each specific function, while the Bayesian component ensures that the selected pdf does not under- or over-fit the available data. Our contribution demonstrates how the prior distribution, necessary in the Bayesian model selection, can be determined, using empirical hyperparameters. That is, the MaxEnt method provides a means to derive a set of possible density functions from various pieces of information and constraints, while Bayesian estimation, being based on empirical hyperparamaters, ensures that a parsimonious model that accounts for the data can be selected. In the event of substantial model uncertainty it may be necessary to average over the models. In this case, a Bayesian approach is necessary to determine the posterior model probabilities conditional on the available data [14, 15]. The approach presented here also addresses the limitation of the method proposed in Hibbert *et al.* [16], by defining an appropriate prior for the Lagrangian parameters.

We have been made aware of a similar approach by Bretthorst [17] which combines both the MaxEnt and Bayesian methods. The approach presented here was independently formulated and developed, and may offer some additional and/or alternative insight into the problem of estimating probability density functions. The difference in the approaches is discussed, in the next section, after some background to the MaxEnt method is given, and our combined MaxEnt and Bayesian approach is further outlined.

1.2 Background

The MaxEnt principle provides a general approach for determining a pdf given the available data and a set of constraints [18, 19]. Maximising the entropy function with respect to a set of constraints ensures that a unique pdf with the fewest assumptions can be determined [5, 19–22]. Shore and Johnson [20, 21] proved that maximising the entropy function with respect to available information is the *only* function that leads to the most probable distribution. They showed the entropy function must satisfy uniqueness, coordinate system invariance, subset and *system independence* [20, 21, 23, 24].

Ishwar and Moulin [22] showed that the MaxEnt method is applicable to a general class of moment inequalities and incorporates a wider class of convex distributions, such as the exponential-family of distributions. There are many examples where specific moment conditions can be used to derive many common density functions, including the exponential, Pareto, normal, Poisson, lognormal and gamma distribution [5, 25–27]. Abramov [28] has shown how the MaxEnt moments approach can be extended to estimate multidimensional pdfs. Importantly, Pressé *et al.* [23, 24] have shown that 'exotic' forms of the entropy function violate the systems independence (i.e. additivity) requirement established by Shore and Johnson [20, 21] and result in pdfs with biases that are not supported by the data. Moreover, Pressé *et al.* [23, 24] argue that pdfs with power laws can be determined by imposing appropriate constraints, rather than by redefining the entropy function.

The MaxEnt moments method proposed by Mead and Papanicolaou [13], determines the underlying density function by maximising the entropy function constrained by a set of polynomial moments. This approach represents a special case of the general approach proposed by Jaynes [18], and the axiomatic argument presented by [20–22]. Mead and Papanicolaou [13] proved that for a monotonic sequence of moments, the underlying density function can be reconstructed to within a selectable error. Mead's and Papanicolaou's MaxEnt moments method is applicable to problems where a monotonic sequence of moments can be used to derive the corresponding density function – see examples in Sec. III to Sec. V in [13].

When limited randomly sampled data are available, unique moments cannot be determined accurately, although the sample moments need to be used in Mead's and Papanicolaou's MaxEnt moments method. Due to the sample size, the sample moments will suffer from sampling uncertainty. In order to address this, Ciulli *et al.* [29] proposed an extension of Mead's and Papanicolaou's method for 'noisy' sample moments. The extension assumed a Gaussian likelihood function to account for the statistical uncertainty of the moments. However, Ciulli *et al.* [29] assumed the statistical uncertainty of the moments converged to a normal distribution by the central limit theorem, which in turn is dependent on having a large sample size. Furthermore, the method can result in under- or over-fitting of the density function when too few or too many moments are used to evaluate the pdf. It does not qualify the appropriate number of moments needed to account for the data. Wu [25] proposed a variety of statistical measures and information criteria to determine the appropriate MaxEnt moment model, and demonstrated its application to U.S. family income, which produced a better fitting result compared to the conventional choices of lognormal and gamma distributions. Park and Bera [27] derived a number of common pdfs using the MaxEnt method, which relies on a set of Lagrangian parameters (see Table 1, p.221), and used orthodox statistical methods to assess competing MaxEnt models. Zellner and Tobias [30] used a Bayesian framework to estimate posterior odds and then select between various regression models.

A key component in Bayesian model selection is the *probabilistic evidence* or marginalized likelihood function, which quantifies how probable the data is conditional on a particular model – for example see [4, 5, 15, 31–38] and references therein. The probabilistic evidence ensures that no particular parameter set, or small region of parameter sets, is overly influential. Once the probabilistic evidence for a given MaxEnt model (i.e. a MaxEnt distribution with specific number of moments) has been determined, the posterior model probability can be evaluated by applying Bayes theorem – see abovementioned references. The posterior model probability quantifies the degree of belief for the MaxEnt pdf model, conditional on the data.

In our solution, we combine the MaxEnt moments method of Mead and Papanicolaou [13] with Bayesian model selection to determine the number of components that best accounts for the data. We treat the Lagrangian multipliers as the model parameters and attribute a prior distribution to them. We use Markov Chain Monte Carlo method (MCMC) to sample the Lagrangian multipliers and calculate the evidence of each given model. MCMC enables the joint pdf of the Lagrangian multipliers and the data to be calculated. The integration over the Lagrangian multipliers to calculate the evidence is accomplished using reverse importance sampling technique – see [32] (Appendix E, pp. 217–222). The priors for the Lagrangian multipliers are defined as a multivariate normal distribution, centered about the optimum set of Lagrangian multipliers for a given model with covariances calculated from the Hessian matrix of the free energy proposed by Mead and Papanicolaou [13]. The optimum set of Lagrangian multipliers and covariance matrix become hyperparameters for the Lagrangian multipliers of a given model, and are evaluated empirically from the data. The uncertainty for the MaxEnt pdf model is calculated using the MCMC sample for the Lagrangian multipliers.

Critically, a multivariate normal distribution was used as the prior pdf for the Lagrangian multipliers. This prior pdf accounts for the cross-correlations between the Lagrangian multipliers, which also impacts on the uncertainty region for the MaxEnt model. This feature defines the difference between our solution and that of Bretthorst [17], which assumed the Lagrangian multipliers are independent and truncated by the minimum and maximum values of the Lagrangian values.

A combined MaxEnt moments and Bayesian approach assesses any features that may appear in the MaxEnt pdfs relative to the data. Furthermore, this approach could be extended to assess a larger set of MaxEnt model structures, where each structure is derived from a variety of constraints, not just from the set of sample moments. Another key observation is that the MaxEnt method is used to derive likelihood functions that are assessed by the Bayesian model selection.

1.3 Paper outline

An outline of the paper is as follows. The MaxEnt and Bayesian theory are presented in Section 2; Section 3 outlines the numerical procedure, which applies a MCMC technique to calculate the evidence and model probabilities. In Sections 4 and 5 the proposed method is tested on simulated datasets, sampled from three known pdfs for a range of sample sizes. The three true pdfs were a Gaussian, non-Gaussian and lognormal, respectively for a range of sample sizes. The uncertainties for the most probable MaxEnt model were also calculated and used to compare the competing MaxEnt model pdfs. Section 5.5 outlines how the present approach can be generalised to selecting competing MaxEnt models derived using moments other than centralised polynomial moments. The conclusion is given in Section 6. Appendix A outlines the MCMC procedure used to calculate the evidence and model probabilities, including pseudo-code of the MCMC algorithm used in the study – see Algorithm 1.

2 Theory

2.1 Transforming the measurand

In our formulation of the problem, centralised moments are used in the MaxEnt method. The data measurand may also otherwise be transformed to a more suitable coordinate system. For example, the transformation may be an offset, to remove the mean, or taking the logarithm, or both. The

measurand is denoted x, and the transformed measurand is denoted $\tilde{x}$.

The MaxEnt pdf is determined in the $\tilde{x}$-coordinate system and then transformed back into the x-coordinate system, where the probabilistic evidence (13), is calculated. This procedure generally avoids numerical instabilities arising from very large moments, such as a singular Hessian matrix, needed for the hyperparameter and defined later in (17).

More specifically, a transformation, denoted $\mathcal{T}$, is found such that the measurand x is transformed into $\tilde{x}$, and is expressed as,

$$\mathcal{T} : x \to \tilde{x} \tag{1a}$$

and

$$\mathcal{T}^{-1} : \tilde{x} \to x. \tag{1b}$$

The final MaxEnt pdf can be expressed in the x-coordinate system by applying the following inverse transformation,

$$f(x) = f(\tilde{x}) \left| \frac{\mathrm{d}\tilde{x}}{\mathrm{d}x} \right|. \tag{2}$$

2.2 MaxEnt moments approach

2.2.1 MaxEnt PDF

In this section, the maximum entropy (MaxEnt) moments method developed by Mead and Papanicolaou [13] is outlined. The details of proofs are given in their paper. The genesis of the approach can be found in Jaynes [18]. The MaxEnt method is used to determine a probability density function from a set of moment constraints. The set of moments represent *testable information* [5, 18, 19]. Subject to this information, the pdf that maximises the entropy function is determined.

In our application of Mead's and Papanicolaou's [13] approach, we transform the problem of determining the pdf according to (1a), and define the first N moments, denoted by $\tilde{\boldsymbol{\mu}} = \{\tilde{\mu}_n;\, n = 1, 2, \ldots, N\}$. The moments are used to constrain the respective MaxEnt pdf (i.e. $f(\tilde{x})$) and are defined as,

$$\tilde{\mu}_n = \int_{\tilde{x}\in\mathbb{R}} \tilde{x}^n \, f(\tilde{x}) \, \mathrm{d}\tilde{x}, \quad \forall\, n = 1, 2, \ldots, N. \tag{3}$$

We wish to find the MaxEnt pdf in the transformed coordinates, $f(\tilde{x})$, for different model structures. The model structure, obtained by constraining the first N moments of $f(\tilde{x})$ to equal $\tilde{\boldsymbol{\mu}}$ is denoted by $\mathcal{M}_N$. For $\mathcal{M}_N$, we determine the $f(\tilde{x})$ that maximises the entropy function subject to the set of N moments. This ensures that we obtain the pdf with the fewest assumptions subject to the constraints. The entropy function of $f(\tilde{x})$ is given by the Boltzmann-Shannon entropy function [5, 18, 19],

$$S\left[f(\tilde{x})|\, m(\tilde{x})\right] \;=\; -\int_{\tilde{x}\in\mathcal{X}} f(\tilde{x}) \, \ln\left[f(\tilde{x})/m(\tilde{x})\,\right] \mathrm{d}\tilde{x}, \tag{4}$$

where $m(\tilde{x})$ plays the role of "invariant measure"[1] for $S\left[f(\tilde{x})|\, m(\tilde{x})\right]$, such that $m(\tilde{x}) \geq 0$, $\forall\, \tilde{x}$; and where $\tilde{x} \in \mathcal{X} \subset \mathbb{R}$, where $\mathcal{X} = [\tilde{x}_{min}, \tilde{x}_{max}]$ and defines the finite boundary for $f(\tilde{x})$. We note that strictly speaking, (4) defines a functional of $f(\tilde{x})$ relative to $m(\tilde{x})$. The measure $m(\tilde{x})$ ensures that (4) is invariant when (1) is applied. The entropy function (4) is also equivalent to minus the Kullback-Leibler divergence, $S[f(\tilde{x})|\, m(\tilde{x})] = -D_{KL}(f(\tilde{x})||\, m(\tilde{x}))$ [4].

[1] Sivia [5] describes $m(\tilde{x})$ as a Lebesgue measure [19, 39].

At a practical level $m(\tilde{x})$ represents the *default* MaxEnt model. That is, a prior estimate, guess, assumption or theoretically derived estimate about $f(\tilde{x})$ prior to collection of the data. Once the data has been collected, the MaxEnt principle updates $m(\tilde{x})$ with respect to a set of constraints – see below (5) and (6). If *a priori* information is available about the measurand, this information can be used to define $m(\tilde{x})$ (using the MaxEnt principle). For example, suppose the possible minimum and maximum range of the data is known *a priori*. According to the MaxEnt principle the appropriate $m(\tilde{x})$ would be a uniform pdf over the range of the data. If additional information in terms of the mean and standard deviation were known *a priori*, then $m(\tilde{x})$ would be defined as a normal pdf via the MaxEnt principle. That is, using the available *a priori* information about the measurand, the MaxEnt principle can be used to express the prior information as a density function in the form of $m(\tilde{x})$. Both $f(\tilde{x})$ and $m(\tilde{x})$ are normalised for unit area.

Following the approach of Mead and Papanicolaou [13], we express the Lagrangian function as,

$$\begin{aligned}\mathcal{L}[f(\tilde{x})] &= -\int_{\mathcal{X}} f(\tilde{x})\,\ln[f(\tilde{x})/m(\tilde{x})]\,\mathrm{d}\tilde{x} + \lambda_0\left[1-\int_{\mathcal{X}} f(\tilde{x})\,\mathrm{d}\tilde{x}\right] \\ &+ \sum_{n=1}^{N}\lambda_n\left[\tilde{\mu}_n - \int_{\mathcal{X}}\tilde{x}^n\,f(\tilde{x})\,\mathrm{d}\tilde{x}\right],\ \forall N, \end{aligned} \tag{5}$$

where λ_0 and $\boldsymbol{\lambda} = \{\lambda_n;\, n = 1,\ldots,N\}$ are the Lagrangian parameters corresponding to the constraints on moments $0, 1, \ldots, N$. The first term in (5) is the entropy function, (4), the second term constrains $f(\tilde{x})$ to have unit area, and the third term is the set of constraints imposed by the sample moments, given later by (19).

Following the procedure of maximising the entropy function with respect to λ and $f(\tilde{x})$, under the moment constraints, we set $\delta\mathcal{L} = 0$ and arrive at [13],

$$f(\tilde{x}) \equiv f(\tilde{x}|\,\boldsymbol{\lambda}_N,\,\mathcal{M}_N,\,\mathcal{I}) \tag{6a}$$

$$= m(\tilde{x})\,\exp\left[-1-\sum_{n=0}^{N}\lambda_n\,\tilde{x}^n\right] \tag{6b}$$

$$= \frac{m(\tilde{x})}{Z(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I})}\,\exp\left[-\sum_{n=1}^{N}\lambda_n\,\tilde{x}^n\right],\ \forall N, \tag{6c}$$

where $\boldsymbol{\lambda_N} = \{\lambda_n;\, n = 1,\ldots,N\}$ is the set of Lagrangian multipliers for model structure $\mathcal{M}_N$; $Z(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I}) = \mathrm{e}^{1+\lambda_0}$ and is the normalising term or *partition function*; and $\mathcal{I}$ represents all the background information that might be available.

The $Z(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I})$ term in (6), also denoted $Z(\boldsymbol{\lambda}_N)$ for brevity, ensures the pdf is normalised for unit area, and is given by

$$Z(\boldsymbol{\lambda}_N) \equiv Z(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I}) \tag{7a}$$

$$= \int_{\mathcal{X}} m(\tilde{x})\,\mathrm{e}^{-\sum_{n=1}^{N}\lambda_n\,\tilde{x}^n}\,\mathrm{d}\tilde{x}. \tag{7b}$$

More importantly, $Z(\boldsymbol{\lambda}_N)$ plays the role of the partition function [18] and ensures that a unique set of Lagrangian multipliers can be determined from moments, $\tilde{\boldsymbol{\mu}}$.

The MaxEnt pdf $f(\tilde{x})$ can also be expressed in discrete form as $\mathbf{f} = \{f_k = f(\tilde{x}_k);\, k = 1, 2, 3, \ldots, K\}$ from (6) for K-sample points of $\tilde{x}$, and where the explicit dependence on $\boldsymbol{\lambda}_N$ has again been omitted.

2.2.2 Determining the Lagrangian multipliers

The uniqueness of the MaxEnt pdf, for a given number of moments, is dependent on uniquely determining $\boldsymbol{\lambda}_N$. An equivalent formulation of a maximum entropy approach is to derive the free energy in terms of $\boldsymbol{\lambda}$ and determine the global minimum.

Mead and Papanicolaou [13] showed that the free energy functional in terms of $Z(\boldsymbol{\lambda})$ is given by,

$$\mathcal{F}(\boldsymbol{\lambda}_N|\,\boldsymbol{\mu},\,\mathcal{M}_N,\,\mathcal{I}) = \ln Z(\boldsymbol{\lambda}_N) + \sum_{n=1}^{N} \tilde{\mu}_n\,\lambda_n, \tag{8a}$$

and is minimized when,

$$\nabla\mathcal{F} = 0, \tag{8b}$$

with the solution to (8b), which minimises (8a), denoted by $\hat{\boldsymbol{\lambda}}_N$.

The convexity of (8a) can be proven by showing that the Hessian matrix of $Z(\boldsymbol{\lambda}_N)$ is positive definite [13]. Mead and Papanicolaou [13] point out that the convexity of (8a) ensures that the Lagrangian multipliers can be determined uniquely. In practice $\hat{\boldsymbol{\lambda}}_N$ can be determined by using nonlinear optimisation algorithms and is discussed further in Sect. 3.2.

Mead and Papanicolaou [13] also proved the uniqueness of the MaxEnt moment pdf for a monotonic sequence of moments [13]. Monotonicity holds if the moments are not polluted by statistical noise or sampling errors. However, the monotonicity of moments begins to break down if estimates of the moments are noisy, which is the case for limited data.

In reality, the moments are noisy, which needs to be taken into account when determining the MaxEnt pdfs, in order to prevent under- and over-fitting of the experimental data. It is this issue which is the motivation for the present research where Bayesian model selection is applied to determine the most probable MaxEnt pdf given the available data and prior information.

2.3 Bayesian model selection

In this section, we outline how Bayesian model selection can be applied to determine the posterior model probabilities for competing MaxEnt solutions.

Bayes' theorem for the posterior probabilities over a set of MaxEnt models $\{\mathcal{M}_N\}, N = 1,\ldots,\mathcal{N}$, conditional on the data $\mathbf{D}$, and background information $\mathcal{I}$, is given by,

$$p(\mathcal{M}_N|\,\mathbf{D},\,\mathcal{I}) = \frac{p(\mathcal{M}_N|\,\mathcal{I})\,p(\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})}{p(\mathbf{D}|\,\mathcal{I})}, \quad \forall\, N = 1,2,3,\ldots,\mathcal{N}, \tag{9}$$

where $p(\mathcal{M}_N|\,\mathbf{D},\,\mathcal{I})$ is the posterior model probability; $p(\mathcal{M}_N|\,\mathcal{I})$ is the prior model probability for the N-moment MaxEnt model; $p(\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})$ is the *probabilistic evidence* or integrated likelihood of the data $\mathbf{D}$, conditional on a given MaxEnt model $\mathcal{M}_N$, and $\mathcal{I}$; and the denominator $p(\mathbf{D}|\,\mathcal{I})$, ensures the probabilities are normalised to unity, where it is implicit in the background information $\mathcal{I}$ that only the specified model structures are being considered.

The *a posteriori* model probability, (9), quantifies the probability that a MaxEnt model, $\mathcal{M}_N$, is correct conditional on the data $\mathbf{D}$, and background assumptions $\mathcal{I}$. In a Bayesian context, a probability represents the *degree of belief* that a proposition is correct, while being conditional on data and background information/assumptions [5, 19]. The posterior model probability is a result of assessing the prior information in the presence of the likelihood function. The prior model probability in (9)

quantifies our belief that each model structure is correct. Hence, a uniform prior model probability asserts the belief that the MaxEnt models in the set are equally probable.

The key component in (9) is the probabilistic evidence, which quantifies the likelihood that the data is a consequence of a given MaxEnt model $\mathcal{M}_N$. This requires integrating over the possible values of the Lagrangian multipliers that define the MaxEnt pdf. Moreover, by integrating over the Lagrangian multipliers, Ockham's razor is embedded. That is, a simple model with a few parameters, that makes moderately successful predictions over a wide domain may be preferable to a complex model, with more parameters, that is highly accurate over a small domain, but otherwise inaccurate [4, 5, 19].

The first step in determining the evidence is understanding the likelihood function of the data in terms of the set of Lagrangian multipliers, $\boldsymbol{\lambda}_N$. The likelihood function for the data $\mathbf{D}$, is the product of MaxEnt pdf evaluations from (6), for a given set of Lagrangian parameters,

$$p(\mathbf{D}|\,\boldsymbol{\lambda}_N,\,\mathcal{M}_N,\,\mathcal{I}) = \prod_{i=1}^{M} f(D_i|\,\boldsymbol{\lambda}_N,\,\mathcal{M}_N,\,\mathcal{I}), \quad \forall\, N, \tag{10}$$

where it is assumed that the data is independently and identically drawn. In order to evaluate (10) the MaxEnt model must be transformed into measurand coordinate system (1b) – Appendix A, Algorithm 1, lines 7 and 8.

Using the likelihood function (10), the joint density function for the data $\mathbf{D}$, and the Lagrangian multipliers $\boldsymbol{\lambda}_N$ is defined as,

$$\begin{aligned} p(\boldsymbol{\lambda}_N,\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I}) &= p(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I})\,p(\mathbf{D}|\,\boldsymbol{\lambda}_N,\,\mathcal{M}_N,\,\mathcal{I}) &\text{(11a)}\\ &= p(\boldsymbol{\lambda}_N|\,\mathbf{D},\,\mathcal{M}_N,\,\mathcal{I})\,p(\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I}), \quad \forall\, N, &\text{(11b)}\end{aligned}$$

where $p(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I})$ is the prior pdf for the Lagrangian multipliers $\boldsymbol{\lambda}_N$, and is conditional on the MaxEnt model $\mathcal{M}_N$. Rearranging (11), the posterior pdf for $\boldsymbol{\lambda}_N$ is given by,

$$p(\boldsymbol{\lambda}_N|\,\mathbf{D},\,\mathcal{M}_N,\,\mathcal{I}) = \frac{p(\boldsymbol{\lambda}_N|\,\mathcal{M}_N,\,\mathcal{I})\,p(\mathbf{D}|\,\boldsymbol{\lambda}_N,\,\mathcal{M}_N,\,\mathcal{I})}{p(\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})}, \quad \forall\, N. \tag{12}$$

The probabilistic evidence introduces Ockham's razor, and can be determined by integration over $\boldsymbol{\lambda}_N$,

$$p(\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I}) = \int_{\mathbb{R}^N} p(\boldsymbol{\lambda}_N,\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})\,\mathrm{d}\boldsymbol{\lambda}_N, \quad \forall\, N. \tag{13}$$

The practical difficulty in applying Bayes theorem for MaxEnt model selection is evaluating (13) for a set of MaxEnt models. This integration of (13) is non-trivial, whether performed analytically and/or numerically.

As noted above, the denominator in (9) ensures that the probabilities are normalised to unity, and requires summing over all the MaxEnt models $\{\mathcal{M}_N;\,\forall\, N\}$,

$$p(\mathbf{D}|\,\mathcal{I}) = \sum_{N=1}^{\mathcal{N}} p(\mathbf{D},\,\mathcal{M}_N|\,\mathcal{I}). \tag{14}$$

In summary, the Bayesian theory outlined in this section addresses the problem of how to determine the posterior model probabilities for a set of MaxEnt solutions by applying Bayes theorem. The next section outlines the challenging task of integrating over the Lagrangian multipliers in order to calculate the probabilistic evidence, (13).

3 Numerical implementation

In this section, the numerical implementation for applying the MaxEnt moments and Bayesian model selection approach, presented in Sec. 2, is outlined.

3.1 MaxEnt moments method

The numerical application of the MaxEnt method required the data to be transformed according to (1a) into $\tilde{\mathbf{x}}$, from which the moments were determined. The MaxEnt density in the transformed space, $f(\tilde{x})$, was then obtained by calculating the Lagrangian parameters by minimising the free energy in (8a) using a nonlinear optimisation algorithm[2]. The MaxEnt model $f(\tilde{x})$ was then inverse transformed according to (1b) to produce the likelihood function in (10), which was then evaluated using the experimental data $\mathbf{D}$.

The integration required in (7b) was approximated with Gaussian quadrature. The quadrature weights and range of integration were specified by $\{\mathbf{w},\, \tilde{\boldsymbol{x}}\} = \{w_k,\, \tilde{x}_k;\, k = 1, \ldots,\, K\}$ over the boundary of $[\tilde{x}_{min},\, \tilde{x}_{max}]$ using $K = 400$ sample points. The integration range, $[\tilde{x}_{min},\, \tilde{x}_{max}]$ was determined from the transformed data and ensured the tails of the MaxEnt pdf extended well beyond the data range.

Using this $\tilde{\boldsymbol{x}}$-range, the MaxEnt pdf was evaluated in the $\tilde{x}$-domain using (6c) with Z in (7) approximated by,

$$Z(\boldsymbol{\lambda}_N|\, \mathcal{M}_N,\, \mathcal{I}) \approx \sum_{k=1}^{K} w_k\, m(\tilde{x}_k)\, \exp\left[-\sum_{n=1}^{N} \lambda_n\, \tilde{x}_k^n\right], \quad \forall\, N. \tag{15}$$

The MaxEnt pdf in the original x-domain was obtained by transforming each quadrature node back into the measurand space using (1b), then interpolating between the quadrature nodes. The likelihood function (10) was evaluated from the resulting interpolated points from this pdf.

3.2 Defining the prior for Lagrangian multipliers

To evaluate the evidence for each model, a prior distribution for each model's parameter set $\boldsymbol{\lambda}_N$ must be specified. We chose to use a Gaussian prior for each $\boldsymbol{\lambda}_N$, with hyperparameters evaluated empirically. The prior was centred around the MaxEnt $\boldsymbol{\lambda}$ solution, $\hat{\boldsymbol{\lambda}}_N$, for each model and the covariance matrix, denoted $\mathbf{C}_N$, was selected so that the prior distribution reflected how sensitive the entropy was to $\boldsymbol{\lambda}_N$ around the MaxEnt pdf solution. Priors obtained by this method are concentrated near feasible solutions and they penalise mismatches between the model and the data due to under- or over-fitted model structures, rather than penalise a poorly restricted prior. For example, using a prior centred around the origin would assign significant prior mass to unstable solutions; or using a prior with hypersphere probability density contours (so that each dimension were equally weighted) would overly penalise higher-order models, by giving significant prior mass to unlikely solutions.

To include the sensitivity of the entropy to $\boldsymbol{\lambda}_N$, we used the free energy, $\mathcal{F}(\boldsymbol{\lambda}_N)$, given by (8a), noting that $\mathcal{F}(\boldsymbol{\lambda}_N)$ can be obtained by evaluating the entropy (4), at solutions given in (6). That is, $\mathcal{F}(\boldsymbol{\lambda}_N) = S[f(\tilde{x}|\, \boldsymbol{\lambda}_N,\, \mathcal{M}_N,\, \mathcal{I})] = \mathbb{E}[-\ln(f(\tilde{x}|\, \boldsymbol{\lambda}_N,\, \mathcal{M}_N,\, \mathcal{I})]$. In particular, we fitted the Gaussian

[2]More specifically, *Mathematica*'s `NMinimize` function and associated package was used to numerically calculate $\hat{\boldsymbol{\lambda}}_N$ from (8a).

prior to the distribution proportional to $\exp\left[-\mathcal{F}(\boldsymbol{\lambda}_N)\right]$, using the second-order Taylor series expansion of $\mathcal{F}(\boldsymbol{\lambda}_N)$ around $\hat{\boldsymbol{\lambda}}_N$. That is,

$$\mathcal{F}(\boldsymbol{\lambda}_N) = \mathcal{F}(\hat{\boldsymbol{\lambda}}_N) + \frac{1}{2}(\boldsymbol{\lambda}_N - \hat{\boldsymbol{\lambda}}_N)^T \nabla\nabla\mathcal{F}(\hat{\boldsymbol{\lambda}}_N)(\boldsymbol{\lambda}_N - \hat{\boldsymbol{\lambda}}_N) + \cdots, \tag{16}$$

since $\nabla\mathcal{F}(\hat{\boldsymbol{\lambda}}_N) = 0$ by (8b). We denote the Hessian of $\mathcal{F}(\boldsymbol{\lambda}_N)$ by $\mathbf{H}_N$, such that $\mathbf{H}_N = \nabla\nabla\mathcal{F}(\hat{\boldsymbol{\lambda}}_N)$. From [40] $\det\mathbf{H}_N > 0$, where $\mathbf{H}_N$ was calculated explicitly for a given model in [40] as,

$$\begin{aligned} H_{nm} &= \frac{\partial^2}{\partial\lambda_n\partial\lambda_m}\mathcal{F}(\boldsymbol{\lambda}_N) && \text{(17a)} \\ &= \langle\tilde{x}^{n+m}\rangle - \langle\tilde{x}^n\rangle\langle\tilde{x}^m\rangle, \quad \forall\, m, n = 1,\, 2,\, 3, \ldots,\, N, && \text{(17b)} \end{aligned}$$

where,

$$\langle\tilde{x}^{n+m}\rangle = \int_{\tilde{x}_{min}}^{\tilde{x}_{max}} \tilde{x}^{n+m}\, f(\tilde{\boldsymbol{x}}|\,\hat{\boldsymbol{\lambda}}_N,\, \mathcal{M}_N,\, \mathcal{I})\,\mathrm{d}\tilde{x}, \tag{17c}$$

and similarly for $\langle\tilde{x}^n\rangle$ and $\langle\tilde{x}^m\rangle$. As such and in general, $\mathbf{H}_N$ is positive definite so that $\mathbf{C}_N = \mathbf{H}_N^{-1}$ and the prior for $\boldsymbol{\lambda}_N$ is,

$$p(\boldsymbol{\lambda}_N|\,\hat{\boldsymbol{\lambda}}_N,\, \mathbf{C}_N,\, \mathcal{M}_N,\, \mathcal{I}) = \det[2\pi\,\mathbf{C}_N]^{-\frac{1}{2}}\, \exp\left[-\frac{1}{2}(\boldsymbol{\lambda}_N - \hat{\boldsymbol{\lambda}}_N)^{\mathsf{T}}\,\mathbf{C}_N^{-1}\,(\boldsymbol{\lambda}_N - \hat{\boldsymbol{\lambda}}_N)\right]. \tag{18}$$

In order to numerically determine the Lagrangian parameters $\hat{\boldsymbol{\lambda}}_N$ in (18), we transform a dataset using (1a), which usually at least includes mean removal, then use the first N moments, denoted $\tilde{\boldsymbol{\mu}} = \{\tilde{\mu}_n;\, n = 1, 2, \ldots,\, N\}$, of the transformed dataset as the moment constraints. The untransformed dataset is assumed an independent and identically distributed sample from a true (unknown) distribution, denoted by $\mathbf{D} = \{D_i;\, i = 1,\, 2,\, 3, \ldots,\, M\}$ and the transformed dataset is denoted $\tilde{\mathbf{x}}$. The moment constraints $\tilde{\mu}_n$ are specified by the sample moments, in the transformed space,

$$\tilde{\mu}_n = \sum_{i=1}^{M} \tilde{x}_i^n / M, \quad \forall\, n = 1, 2, \ldots, N, \tag{19}$$

and are used in (8) to determined $\hat{\boldsymbol{\lambda}}_N$. As noted in Sect. 2.2.2 this requires minimising (8a) using nonlinear optimisation. In the present study *Mathematica*'s nonlinear optimisation package was used calculate the Lagrangian parameters (also see footnote 2).

Since the hyperparameters were calculated from the data, the proposed Bayesian model selection is an empirical Bayesian approach, whereas a strict Bayesian approach would require defining an appropriate hyperprior pdf for $\hat{\boldsymbol{\lambda}}_N$ and $\mathbf{C}_N$ and integrating over its domain, using such techniques as Gibbs sampling [3].

The aforementioned discussion concerning the priors for Lagrangian multipliers assumes the set of MaxEnt models were derived using the same transformation of the measurand (1a). Hence, the calculation of the joint-posterior pdf (11a) and the subsequent posterior model probability (9) for each model in the set can be calculated and from which the most probable MaxEnt models can be chosen. If MaxEnt models from multiple sets were to be assessed against each other as a single set of models, where each set of MaxEnt models were derived from different measurand transformations (1a), then the prior for the Lagrangian multipliers (18) from each set would need to be transformed according to (1b). This would enable the posterior model probabilities for all the models to be compared and ranked against each other.

3.3 Model prior

The posterior model probabilities (9), were calculated assuming a uniform pdf for the model prior pdf, $p(\mathcal{M}_N|\mathcal{I})$. This choice of pdf over the range of MaxEnt models assumed the models were equally probable. If *a priori* information about the models were available, it could have been used to define $p(\mathcal{M}_N|\mathcal{I})$. For the set of MaxEnt models, $N = 2, 3, \cdots, \mathcal{N}$, the posterior model probabilities were ranked in a consistent manner and the most probable MaxEnt model determined. In principle, the posterior model probabilities can be used to determine the weighted-average MaxEnt model (e.g. see [14]).

3.4 Outline of the Markov Chain Monte Carlo method

3.4.1 MCMC/Metropolis algorithm

Calculating the probabilistic evidence (13), for the MaxEnt models is a challenging computational task, since it requires integrating over the Lagrangian multipliers. A Markov Chain Monte Carlo (MCMC) method based on a Metropolis algorithm was used to sample $\boldsymbol{\lambda}_N$-space for a given model in proportion to the joint-pdf of (11). After serial correlations were removed, the sampled Lagrangian parameters were then used to compute the model's probabilistic evidence (13), by applying reverse importance sampling algorithm (e.g. see [32]).

The MCMC/Metropolis algorithm is described in Appendix A, including the steps to tune the algorithm and remove the serial correlations from the MCMC data. The algorithm generates a random sequence of $\boldsymbol{\lambda}_N$-values based on the acceptance/rejection of a *trial* $\boldsymbol{\lambda}_N$, where the *acceptance probability* is defined

$$\alpha(\boldsymbol{\lambda}_{(l)},\, \boldsymbol{\lambda}_{trial}) = \min\left(1,\, \frac{p(\boldsymbol{\lambda}_{trial},\, \mathbf{D}|\, \mathcal{M}_N,\, \mathcal{I})}{p(\boldsymbol{\lambda}_{(l)},\, \mathbf{D}|\, \mathcal{M}_N,\, \mathcal{I})}\right). \tag{20}$$

Equation (20) determines the probability of accepting the $\boldsymbol{\lambda}_{trial}$ with respects to the current $\boldsymbol{\lambda}_{(l)}$; if $\alpha \geq 1$ then $\boldsymbol{\lambda}_{trial}$ is accepted unconditionally. For $0 < \alpha < 1$, $\boldsymbol{\lambda}_{trial}$ is only accepted if $\alpha > r$, otherwise rejected, where r is a real random number drawn from a uniform distribution over [0, 1] – see lines 16 to 23 from Appendix A. It can be shown that as $l \to \infty$ the sample of accepted $\boldsymbol{\lambda}_N$-values will be distributed according to (11) from which (12) can be determined (see [32, 35, 41]). The trial Lagrangian multipliers $\boldsymbol{\lambda}_{trial}$, were drawn from a *proposal distribution* and is defined by a multivariate Gaussian density $\boldsymbol{\lambda}_{trial} \sim Normal(\boldsymbol{\lambda}_{(l)},\, \mathbf{C}_{pd})$. The covariance for the proposal density $\mathbf{C}_{pd}$ was defined by $\mathbf{C}_{pd} = c\,\mathbf{C}_N$ with tuning parameter $c > 0$ and $\mathbf{C}_N$ given by the inverse of (17). The tuning involved a set of pre-MCMC simulations from which the optimal acceptance ratio was determined, which in turn determined c.

3.4.2 Numerically calculation of the probabilistic evidence

The serial correlations in the MCMC sample were removed by re-sampling the MCMC data at its autocorrelation length. Using the re-sampled MCMC data, the probabilistic evidence was finally calculated by applying the reverse importance sampling technique [31, 32, 42, 43],

$$p(\mathbf{D}|\, \mathcal{M}_N,\, \mathcal{I}) \approx \left[\frac{1}{Q_r}\sum_{l=1}^{Q_r} \frac{g(\boldsymbol{\lambda}_{(l)N}|\, \bar{\boldsymbol{\lambda}}_N,\, \mathbf{Q}_N)}{p(\boldsymbol{\lambda}_{(l)N},\, \mathbf{D}|\, \mathcal{M}_N,\, \mathcal{I})}\right]^{-1}, \tag{21}$$

where $\lambda_{(l)N}$ denotes the lth MCMC sample (after serial correlations were removed) of λ for the Nth model; Q_r is the total number of MCMC samples remaining after the serial correlations were removed; $g(\boldsymbol{\lambda}_{(l)N}|\,\bar{\boldsymbol{\lambda}}_N,\,\mathbf{Q}_N)$ is an N-dimensional multivariate sampling density function, chosen such that $\int g(\boldsymbol{\lambda}_{(l)N}|\,\bar{\boldsymbol{\lambda}}_N,\,\mathbf{Q}_N)\,\mathrm{d}\boldsymbol{\lambda}_N = 1$, given by,

$$g(\boldsymbol{\lambda}_{(l)N}|\,\bar{\boldsymbol{\lambda}}_N,\,\mathbf{Q}_N) \;=\; \det\left[2\pi\,\mathbf{Q}_N\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\,(\boldsymbol{\lambda}_{(l)N}-\bar{\boldsymbol{\lambda}}_N)^{\mathsf{T}}\mathbf{Q}_N^{-1}(\boldsymbol{\lambda}_{(l)N}-\bar{\boldsymbol{\lambda}}_N)\right], \tag{22a}$$

where $\bar{\boldsymbol{\lambda}}_N$ is the MCMC sample mean,

$$\bar{\boldsymbol{\lambda}}_N = \frac{1}{Q_r}\sum_{l=1}^{Q_r}\boldsymbol{\lambda}_{(l)N} \tag{22b}$$

and the MCMC covariance matrix $\mathbf{Q}_N$, is given by,

$$\mathbf{Q}_N = \frac{1}{Q_r-1}\sum_{l=1}^{Q_r}(\boldsymbol{\lambda}_{(l)N}-\bar{\boldsymbol{\lambda}}_N)^{\mathsf{T}}(\boldsymbol{\lambda}_{(l)N}-\bar{\boldsymbol{\lambda}}_N). \tag{22c}$$

For large Q_r, $\bar{\boldsymbol{\lambda}}_N \approx \hat{\boldsymbol{\lambda}}_N$ and $\mathbf{Q}_N \approx \mathbf{C}_N$, where $\hat{\boldsymbol{\lambda}}_N$ and $\mathbf{C}_N$ are described by (8) and (17), respectively.

Equation (21) is a weighted harmonic mean of the evidence. The role of $g(\boldsymbol{\lambda}_{(l)N}|\,\bar{\boldsymbol{\lambda}}_N,\,\mathbf{Q}_N)$ is to weight the $p(\boldsymbol{\lambda}_{(l)},\,\mathbf{D}|\,\mathcal{M}_N,\mathcal{I})$ such that outliers do not dominate the result. The advantage of using (21), as opposed to the usual importance sampling (IS) techniques, is that $\boldsymbol{\lambda}_N$ can be drawn from the joint-posterior pdf (11) and used to calculate the evidence value. The usual IS techniques assumes *a priori* the posterior pdf for $\boldsymbol{\lambda}_N$ can be approximated by a sampling density [32].

Since $g(\boldsymbol{\lambda}_{(l)N}|\,\bar{\boldsymbol{\lambda}}_N,\,\mathbf{Q}_N)$ is the numerator within the summation of (21), it should have thinner tails for numerical stability reasons, and since the sample size is large a multivariate Gaussian density function is sufficient [32].

Finally it is worth noting that other numerical techniques for calculating the probabilistic evidence (13) could be applied, such as Nested Sampling [36, 44], and Tempering [45]. However, the rationale of using an MCMC algorithm and then applying (21) is that $f(\tilde{\mathbf{x}}|\,\boldsymbol{\lambda}_N,\,\mathcal{M}_N,\mathcal{I})$ can also be sampled within the same algorithm. Although, post-processing of the MCMC sample is required to remove the serial correlations in the Markov chain.

4 Simulated Datasets

In this section an outline of the simulated datasets used to test the MaxEnt/Bayesian method is provided.

4.1 Test distributions

The testing of the MaxEnt/Bayesian approach focused on assessing the reliability of the Bayesian model selection for different types of distributions and data sample sizes, ranging from negatively to positively skewed distributions. Three distribution types were chosen to be a Gaussian (symmetric), non-Gaussian derived from the Lagrangian multipliers (left-skewed) and lognormal (right-skewed) distributions. Table 1 summarises the parameters that defined the test distributions.

Simulated datasets were randomly drawn from the pre-defined distributions given in Table 1 with sample sizes according to Table 2. For a given distribution and sample size, the simulated data were *independently and identically distributed* (iid).

Type	Distribution	Parameters
1	Gaussian	$\langle x \rangle = 30.000$, $\sigma = 3.162$
		$\{\lambda_1, \lambda_2\} = \{0, 0.05\}$
2	non-Gaussian	$\langle x \rangle = 41.000$, $\sigma = 1.476$
		$\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{-0.05, 0.20, 0.08, 0.01\}$
3	Lognormal	$\langle x \rangle = 20.000$, $\sigma = 6.000$
		$(\mu = 2.953, \sigma_0 = 0.294)$

Table 1: Different distribution types used to test the MaxEnt/Bayesian approach. The mean, $\langle x \rangle$, and standard deviation, σ, of the measurand have arbitrary units (au). In the case of the lognormal, μ and σ_0 correspond to the lognormal mean and standard deviation, respectively.

Test	Sample Size, M
1	515
2	250
3	125
4	60
5	30

Table 2: Samples sizes used to generate simulated datasets.

4.2 Simulating datasets

The proposed method was then applied to determine the most probable MaxEnt model for each dataset. These sets of tests also served to indicate the minimum sample size for which the Bayesian model selection could determine the correct model.

For the Gaussian and Lognormal distributions from Table 1, *iid* samples were drawn using standard random sampling algorithms that are available in *R* or *Mathematica*.[3] In the case of the non-Gaussian test distribution the *iid* samples were drawn by numerically evaluating and interpolating the inverse cdf of the test distribution. In particular, for each sample size given by Table 2, an *iid* sample of real numbers was uniformly sampled over the region [0, 1]. The corresponding distribution values were then calculated by interpolating the inverse cdf of the non-Gaussian test distribution to each sampled point. The interpolation was carried out over a broad range to ensure the inverse cdf spanned the range of [0, 1] – e.g., see Figure 4(a).

5 Numerical Results & Discussion

In this section we present the results of applying the MaxEnt/Bayesian method to random samples (simulated data) drawn from a known distribution, described in Table 1. The aim of the tests was to examine the ability of the MaxEnt/Bayesian method to predict the correct model structure, and reconstruct the true pdf, given the available data. The MaxEnt/Bayesian results for the three distributions described in Sec. 4 are presented in Sec. 5.1 to 5.3. In each case, the MaxEnt models were

[3] *Mathematica*'s `RandomVariate` function was used to *iid* from the Gaussian and Lognormal test distributions.

calculated and the MCMC calculations of the posterior model probability were carried out for models consisting of $n = 2, 3, 4, 5$ and 6 moments, from which the most probable model was selected – see Sec. 3 and Appendix A. In Sec. 5.4, the goodness of fit of the MaxEnt pdfs is quantified and compared to the true pdfs for a set of sample sizes.

5.1 MaxEnt Result - Gaussian test

The MaxEnt/Bayesian method was tested using data that was randomly drawn from the Gaussian distribution given in Table 1 with sample sizes from Table 2. The Gaussian distribution provides a simple example in which to test and examine the properties of the proposed method.

The MaxEnt pdfs for tests 1 and 5 are presented in Figures 1 and 2, where the MaxEnt models are compared with the true Gaussian pdf. The log-evidence values and posterior model probabilities for all sample sizes are presented in Table 3 and Figure 3, respectively.

The simulated data were de-meaned before the centralised moments (19) were determined as

$$\tilde{\mathbf{x}} = \mathbf{D} - \langle D \rangle, \tag{23}$$

where $\langle D \rangle$ is the sample mean of the data; and $\tilde{\mathbf{x}} = \{\tilde{x}_i; \forall\, i = 1, \ldots, M\}$. The transformed measurand and Gaussian quadrature $\{\tilde{\mathbf{x}}, \boldsymbol{\omega}\}$ were defined over the range ± 20 with $K = 400$ sample points. The MaxEnt models for $N = 2$ to 6 moments were calculated and transformed back into the measurand $\mathbf{x}$-range.

For each MaxEnt model, the default or prior pdf $m(\tilde{x})$ was assumed to be a uniform pdf over the defined range, and is shown in Figures 1 and 2. From Figures 1 and 2, it is evident that there is information in the data, since the final MaxEnt pdfs are non-uniform distributions with respect to $m(\tilde{x})$ and indicate that their final pdf's have diverged some way from $m(\tilde{x})$. Otherwise, we would expect $f(\tilde{x}|\, \boldsymbol{\lambda}, \mathcal{M}_N, \mathcal{I}) \approx m(\tilde{x})$.

More specifically, this represents the influence the constraints have on the entropy function. The constraints act to increase the statistical divergence between the $f(\tilde{x})$ and $m(\tilde{x})$. In the case of no constraints, the divergence between the $f(\tilde{x})$ and $m(\tilde{x})$ is zero, and $f(\tilde{x}|\, \boldsymbol{\lambda}, \mathcal{M}_N, \mathcal{I}) = m(\tilde{x})$.

From Figure 1(a), with sample size = 515, all the MaxEnt models appear to account for the data. Figure 1(b) provides more detail of the tails of the MaxEnt models. The 99% credibility region for this solution was calculated from the 0.5% and 99.5% quantiles from the MCMC sample described in Appendix A.

For all MaxEnt models, with the exception of the $N = 5$ model, the tails of the models are stable. That is, the tails decrease for increasing $|x|$ and monotonic properties are maintained. The $N = 5$ model is unstable as the tails increase for $x \lesssim 12$ due to the $\lambda_5\, x^5$ term in (6), which is caused by the model trying to fit the isolated data point at 17 a.u and the finite $\tilde{\mathbf{x}}$-range, used for computational practicality.

The MaxEnt pdf for sample size 30 is given in Figure 2(a). From visual inspection, the $N = 6$ model differs from the other models in that it attempts to describe undulations in the data, giving the impression of a multi-modal pdf. The unstable tails of $N = 4$ and 5 models are also evident – see Figure 2(b). The $N = 5$ model has $\lambda_5 = -7.05313 \times 10^{-8}$, which means it essentially mimics the $N = 4$ the model. The $N = 2$ and 3 models appear to produce pdfs similar to the true Gaussian pdf.

Also shown in Figures 1 and 2 are the uncertainty regions for the $N = 2$ model. It is clear that as the sample size decreases, the uncertainty region will increase, while precluding particular solutions.

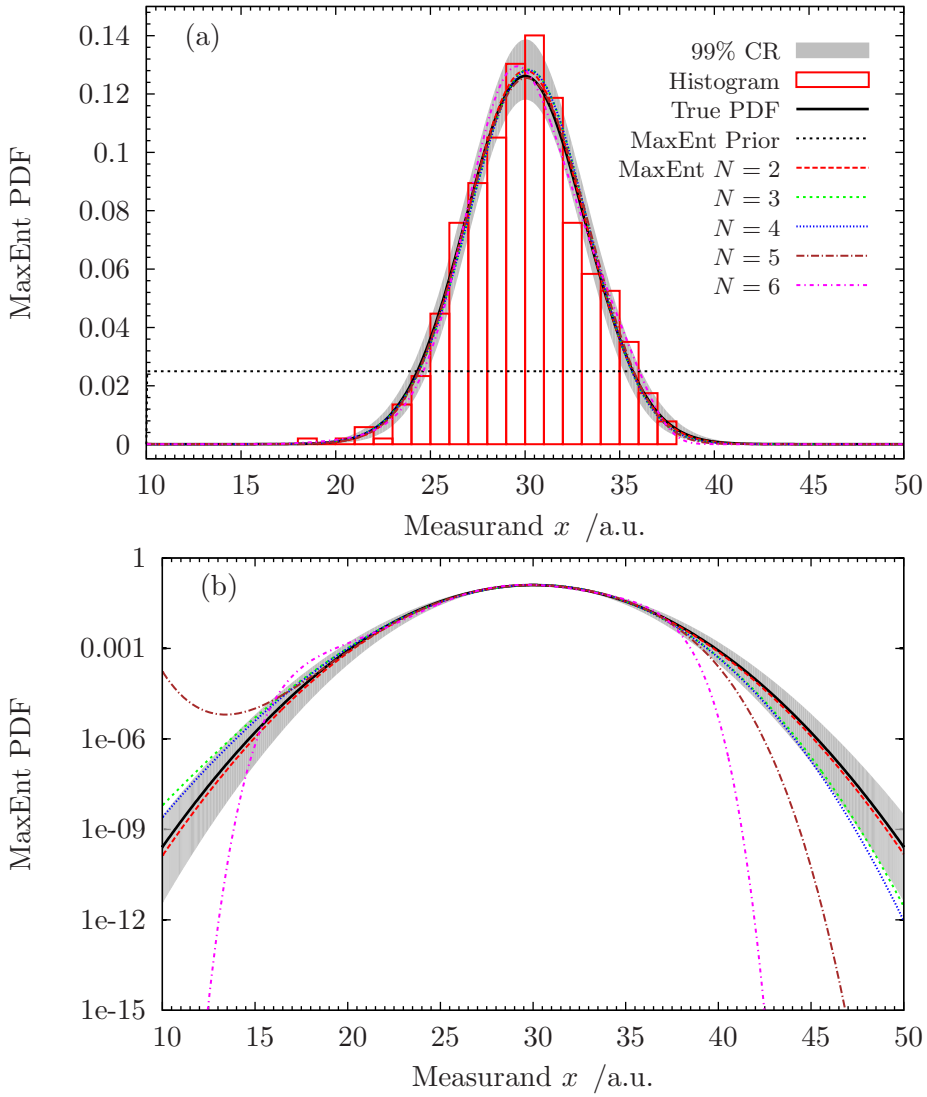

Figure 1: MaxEnt models fitted to dataset of 515 samples, sampled from a Gaussian pdf: (a) Comparing the MaxEnt models with histogram data, prior $m(x)$ and true pdf over a linear-scale. (b) Comparing MaxEnt models and true pdf over a logarithmic scale. The grey shaded area corresponds to the 99% credibility region for the $N = 2$ model (most probable model). All pdfs have been normalised to unit area.

The probabilistic evidence and posterior model probabilities are presented in Table 3 and Figure 3. The probabilistic evidence was calculated using the MCMC procedure outlined in Sec. 3.4 and Appendix A. For all sample sizes considered, the posterior model probabilities predicted the $N = 2$ model as the most probable model. For a given sample size, there is also a systematic dependence on N; that is, the probabilistic evidence decreases as N increases. Given the unstable tails in the MaxEnt models it would be tempting to eliminate them from the model selection, without calculating their probabilistic evidence values. However, this would pre-determine the result by appealing to a heuristic argument, and importantly, if we are going to apply Bayes theorem, all models which are members of the set should be examined and quantified in terms of a probability conditional on the data.

In a MaxEnt context, the $N = 2$ model corresponds to a Gaussian distribution; in a Bayesian context, it is the simplest model that accounts for the data. The probabilistic evidence (13), and subsequent Bayesian probabilities (9), are able to discern between "simple" and "complex" models, even in the case of small sample sizes. This is an important result as the $N = 4$ and 5 models have

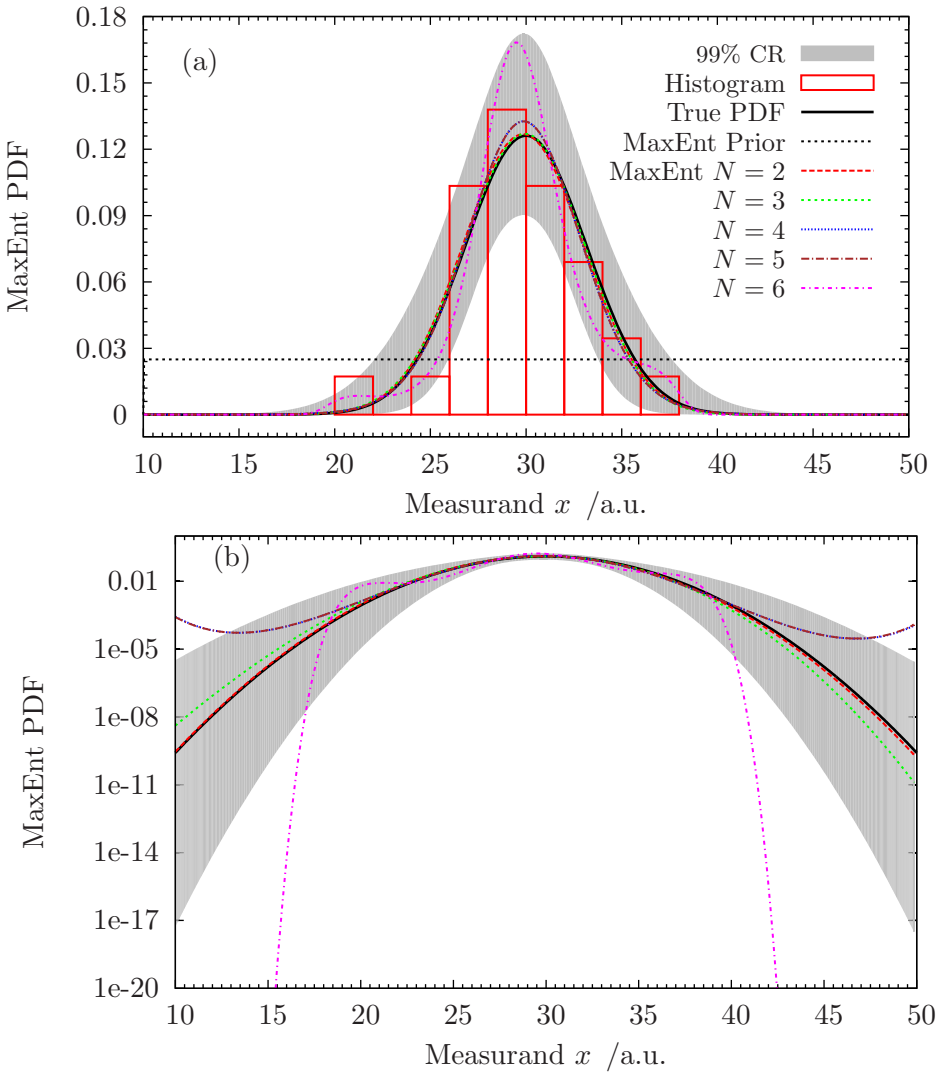

Figure 2: MaxEnt models fitted to dataset of 30 samples, sampled from a Gaussian pdf: (a) Comparing the MaxEnt models with histogram data, prior $m(x)$ and true pdf over a linear-scale. (b) Comparing MaxEnt models and true pdf over a logarithmic scale. The grey shaded area corresponds to the 99% credibility region for the $N = 2$ model (most probable model). All pdfs have been normalised to unit area.

Model	Samples Size=515	250	125	60	30
2	-1322.0	-635.3	-332.3	-160.1	-80.4
3	-1324.8	-637.5	-334.5	-161.7	-82.3
4	-1327.9	-640.2	-336.5	-163.2	-82.5
5	-1329.7	-642.4	-339.3	-163.7	-83.7
6	-1331.0	-644.6	-340.9	-165.0	-85.1

Table 3: Log-probabilistic evidence for Gaussian test over a range of sample sizes. The log-evidence values have been rounded to the first decimal place for convenience, but the maximum relative uncertainty was $\lesssim 0.07\%$ at the 95% confidence level.

undesirable and unphysical features i.e. unstable tails, while the $N = 6$ model over-fits the data. The credibility region at the 99% probability level for $N = 2$ model are also shown in Figures 1 and 2. The credibility region reflects the uncertainty arising from the available sample size and confirms the Bayesian model probabilities. In this case, model averaging does not work because some models are unstable ($N = 4$ and $N = 5$) and a linear weighted instability is still unstable.

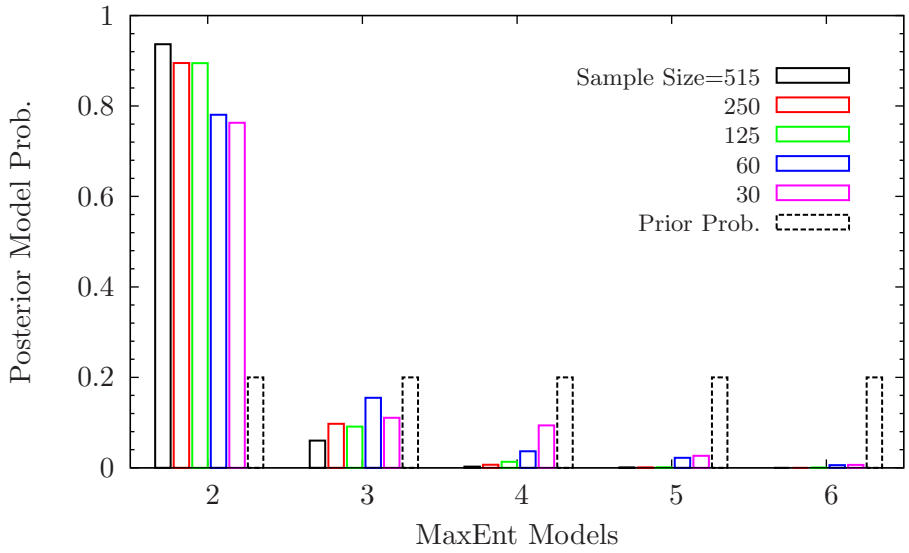

Figure 3: Posterior model probabilities for Gaussian test. MaxEnt models 2 to 6 for a range of sample sizes, 30, 60, 125, 250 and 515. The true distribution was set to a Gaussian pdf (a). In all cases the correct MaxEnt model, $N = 2$ was selected.

The results from the simple example demonstrate the abilities of the MaxEnt method to propose models given the available data, and the Bayesian analysis to "weed out" models that do not account for the data.

5.2 MaxEnt Result - Non-Gaussian test

The second test for the MaxEnt/Bayesian method was based on data (iid) randomly drawn from a pdf which had negative skewness. The pdf was "invented" from a MaxEnt pdf with known Lagrangian multipliers – see Table 1. Figures 4 and 5 show the MaxEnt models for samples sizes of 515 and 60, respectively. The probabilistic evidence values and posterior model probabilities are presented in Table 4 and Figure 6 for all sample sizes.

For this example, the data was transformed according to (23), i.e. demeaned, then centralised sample moments were calculated and used to generate the MaxEnt pdfs for $N = 2$ to 6. By comparing the MaxEnt models to the histogram data in Figure 4(a), it is clear that the simpler MaxEnt models, $N = 2$ and 3 do not (visually) describe the data.

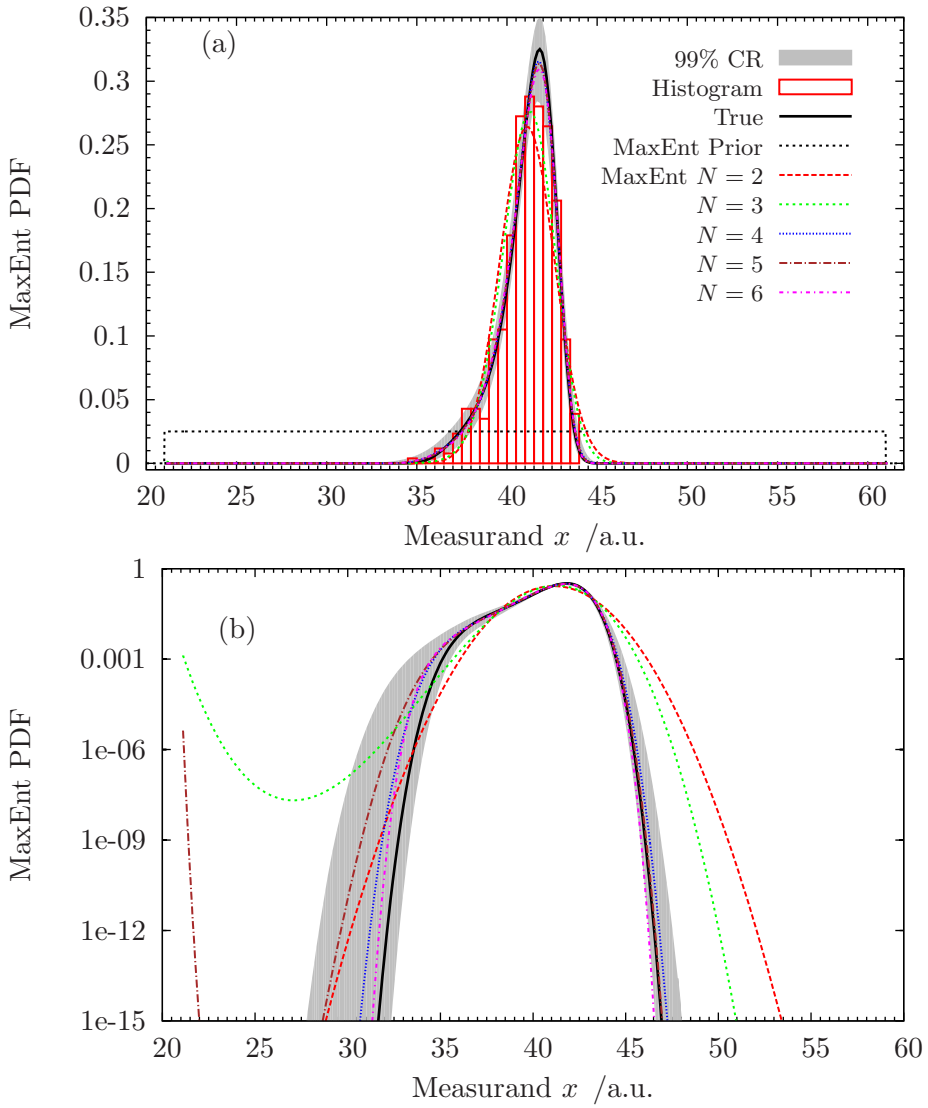

Figure 4: MaxEnt models from dataset of 515 samples, assuming a non-Gaussian pdf: (a) Comparing the MaxEnt models with histogram data, prior $m(x)$ and true pdf over a linear-scale. (b) Comparing MaxEnt models and true pdf over a logarithmic scale. The grey shaded area corresponds to the 99% credibility region for the $N = 4$ model (most probable model). All pdfs have been normalised to unit area.

Figure 5 shows the MaxEnt models from sample size of 60. For these results there is considerable variation between the MaxEnt models due to trying to fit the randomly created high-order feature around 38 au created in the small sample. For example in Figure 5(a), the $N = 6$ model over-fits the data, describing it as a bimodal distribution. The unstable tail for the $N = 3$ model is also evident in Figures 5(a) & (b).

The posterior model probabilities for this case are given in Figure 6, with the probabilistic evidence for the specified sample sizes given in Table 4. From Figure 6 the most probable model selected is the $N = 4$ model for sample sizes of 60 to 515. However, for a sample size of 30 the $N = 2$ model is predicted – also see Table 4. For this case, there is a breakdown of the model selection, indicating that there was inadequate information in the sample for the MaxEnt/Bayesian method to select a

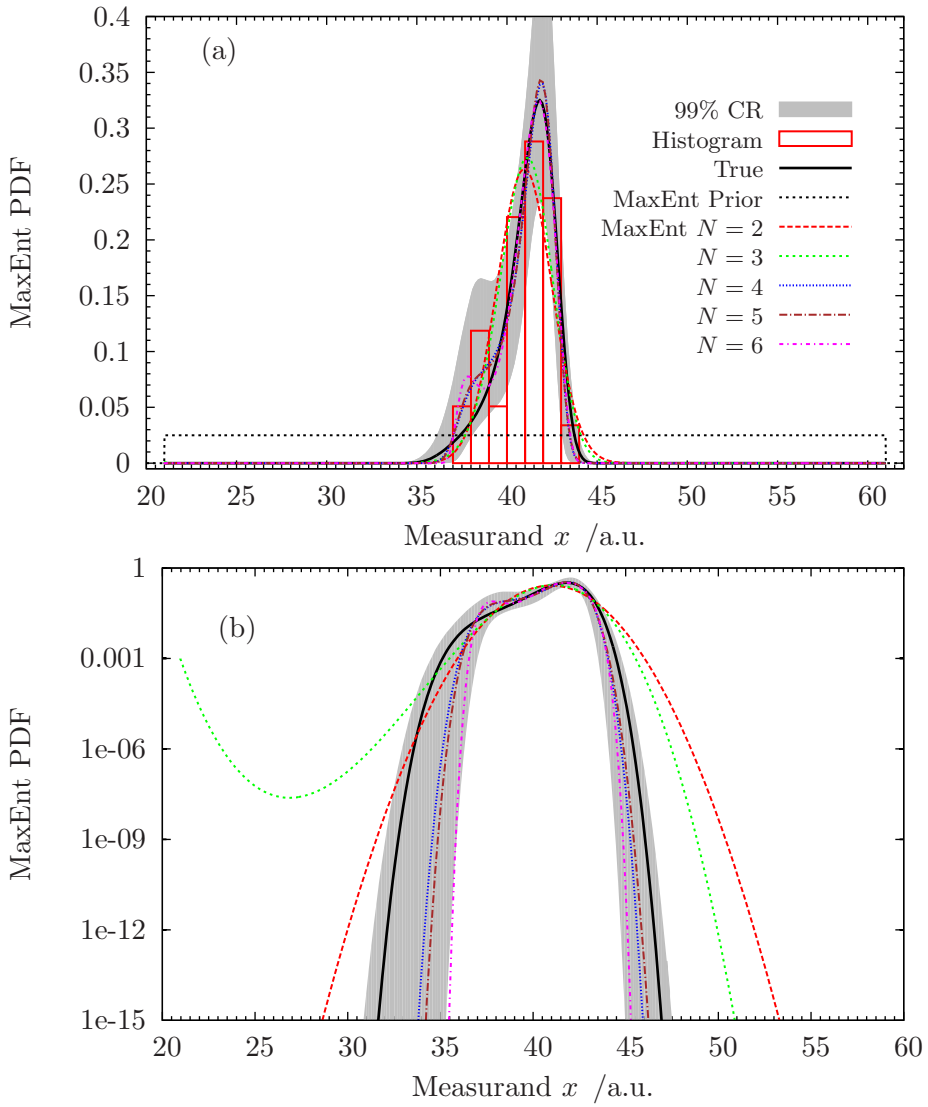

Figure 5: MaxEnt models from dataset of 60 samples, assuming a non-Gaussian pdf (b): (a) Comparing the MaxEnt models with histogram data, prior $m(x)$ and true pdf over a linear-scale. (b) Comparing MaxEnt models and true pdf over a logarithmic scale. The grey shaded area corresponds to the 99% credibility region for the $N = 4$ model (most probable model). All pdfs have been normalised to unit area.

fourth-order distribution.

Model	Sample Size=515	250	125	60	30
2	-949.0	-445.9	-227.7	-114.1	-49.3
3	-935.5	-442.7	-225.7	-113.7	-50.1
4	-915.1	-436.0	-217.4	-111.1	-52.0
5	-916.4	-439.9	-217.8	-111.6	-52.7
6	-921.0	-440.9	-221.2	-114.5	-55.0

Table 4: Log-probabilistic evidence for the non-Gaussian test over a range of sample sizes. The log-evidence values have been rounded to the first decimal place for convenience, but the relative uncertainty is $\lesssim 0.03\%$ at the 95% confidence level.

From Table 4 it is interesting to notice the relative odds between models. For example, the odds that the $N = 4$ model is correct relative to the $N = 3$ model ranges from $e^{20.4} \sim 7 \times 10^8$ to $e^{2.6} \approx 13$ from a sample size of 515 to 60, respectively. In contrast, the odds for $N = 5$ model being correct over $N = 4$ model ranges from $e^{-1.3} \approx 0.27$ to $e^{-0.5} \approx 0.6$, over the same range of sample sizes. We see in the first example there is very strong probabilistic evidence to believe that $N = 4$ model is more probable compares to $N = 3$ model. In the second example, the odds in favor of $N = 4$ model being correct relative to $N = 5$ model has decreased considerably. The $N = 5$ model is mimicking the $N = 4$ model since the $\lambda_5\, x^5$ term is small in comparison to the $\lambda_4\, x^4$ term, and the probabilistic evidence quantifies the *cost* of including an additional parameter in the model.

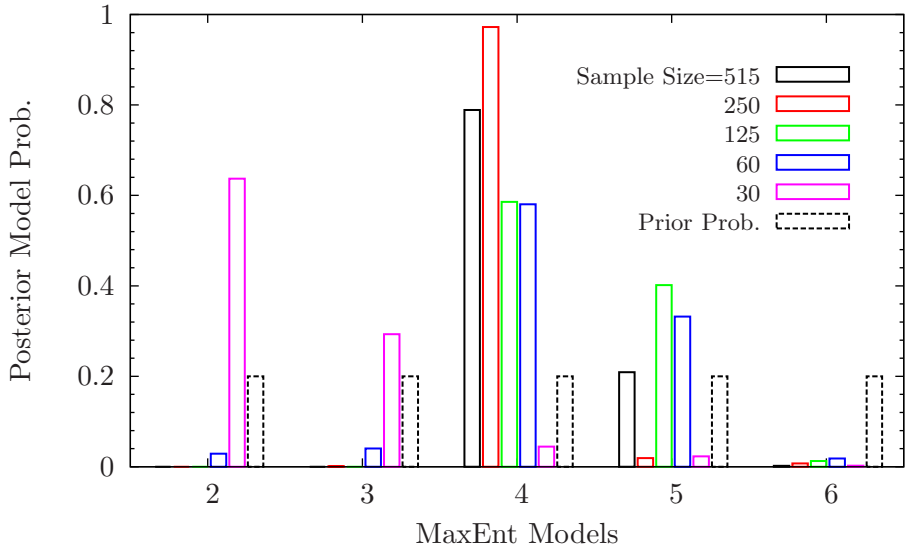

Figure 6: Posterior model probabilities for non-Gaussian test. MaxEnt models 2 to 6 for a range of sample sizes, 30, 60, 125, 250 and 515. The true distribution was set to a Gaussian pdf (b). In all cases the correct MaxEnt model, $N = 4$ was selected.

For this case, the MaxEnt/Bayesian method is able to distinguish between simple and complex models for sample sizes $\gtrsim 60$. It also demonstrates how minimal data can give rise to an incorrect model being predicted. It is not surprising that a lower-order model is selected for a small dataset, because in this case it is more likely that the sample is insufficiently densely populated to represent high-order features and penalise them being omitted from the model.

5.3 MaxEnt Result - Lognormal test

The MaxEnt models shown in Figures 7 & 8 demonstrate how apparently complex models can be transformed and shown to be a simpler model.

In this case the measurand data, **D**, was transformed by

$$\tilde{\mathbf{D}} = \ln \mathbf{D} - \langle \ln \mathbf{D} \rangle, \tag{24}$$

where $\ln \mathbf{D} = \{\ln D_i;\, i = 1, \ldots, M\}$. The corresponding transformed coordinate, $\tilde{\mathbf{x}}$, was defined over $[-4.5,\, 1.5]$ with 400 sample points. The prior MaxEnt model, $m(\tilde{x})$, was defined as a uniform pdf over the same range. The lognormal test-distribution transforms into a normal or Gaussian pdf, and the corresponding MaxEnt distribution is dependent on two moments.

Figure 7 & 8 shows the MaxEnt models for a range of moments and sample sizes. The posterior model probabilities and probabilistic evidence values are also given in Figure 9 and Table 5, respectively. The prior MaxEnt model is also shown in Figures 7 & 8, when transformed back into the measurand coordinates, using (1b) whereupon it changes from a uniform pdf to a Jeffreys prior. That is, $m(x) \propto 1/x$ over the defined range.

In Figure 7(a), there appears to be little variation between all the MaxEnt models. However, on the log-scale Figure 7(b), differences in the tails are revealed; for models $N = 4$ & 5 in the region $x \gtrsim 60$ au the tails are concave up, indicating instability.

Figure 8 again shows greater variation between the MaxEnt models for a sample size of 30. It also suggests greater instability in the models, in particular $N = 3$ & 5 models, while the $N = 6$ model exhibits the similar characteristics to the $N = 6$ model in Figure 7.

Heuristic arguments could be made to eliminate particular MaxEnt models based on the properties of the pdf and tails. However, rather than resorting to such arguments, the probability for a particular model being true, conditional on the data, provides a means to discriminate between competing models. The model probabilities and probabilistic evidence values, presented in Figure 9 and Table 5, show strong and consistent evidence conditional on the data for the $N = 2$ model.

Model	Samples Size=515	250	125	60	30
2	-1648.3	-785.3	-391.6	-201.7	-96.5
3	-1651.2	-786.2	-394.3	-204.0	-97.8
4	-1652.4	-789.8	-395.6	-205.7	-99.0
5	-1657.6	-790.5	-396.4	-208.3	-99.6
6	-1657.0	-793.1	-398.7	-206.9	-101.8

Table 5: Log-probabilistic evidence for lognormal test over a range of sample sizes. The log-evidence values have been rounded to the first decimal place for convenience, but the relative uncertainty is $\lesssim 0.03\%$ at the 95% confidence level.

5.4 Goodness of fit

The model selection process assesses each model structure (i.e. each number of moments used in the model) by averaging the likelihood of the parameters over the prior distribution for the parameter. We are also interested in how well the particular MaxEnt pdf for each model structure matches the true pdf.

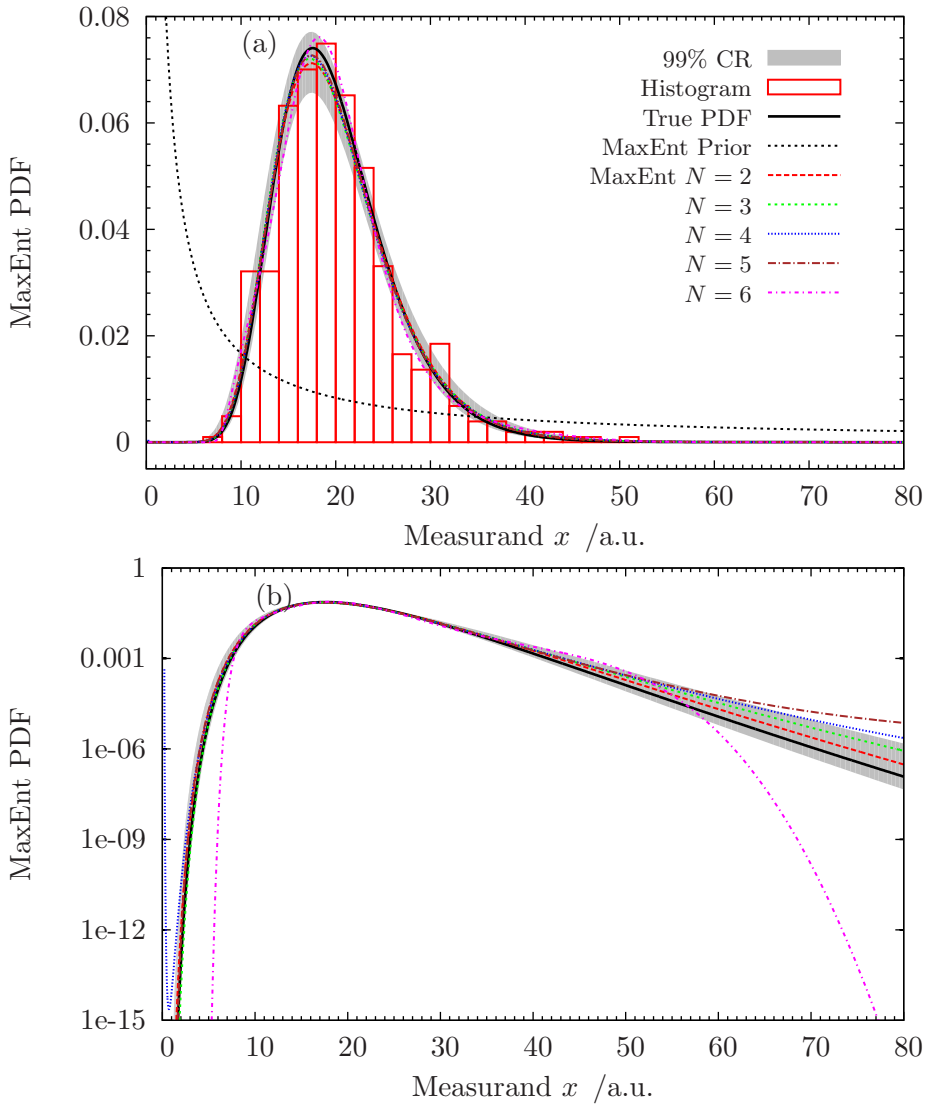

Figure 7: MaxEnt models from dataset of 515 samples, assuming a lognormal pdf: (a) Comparing the MaxEnt models with histogram data, prior $m(x)$ and true pdf over a linear-scale. (b) Comparing MaxEnt models and true pdf over a logarithmic scale. The grey shaded area corresponds to the 99% credibility region for the $N = 2$ model (most probable model). All pdfs have been normalised to unit area.

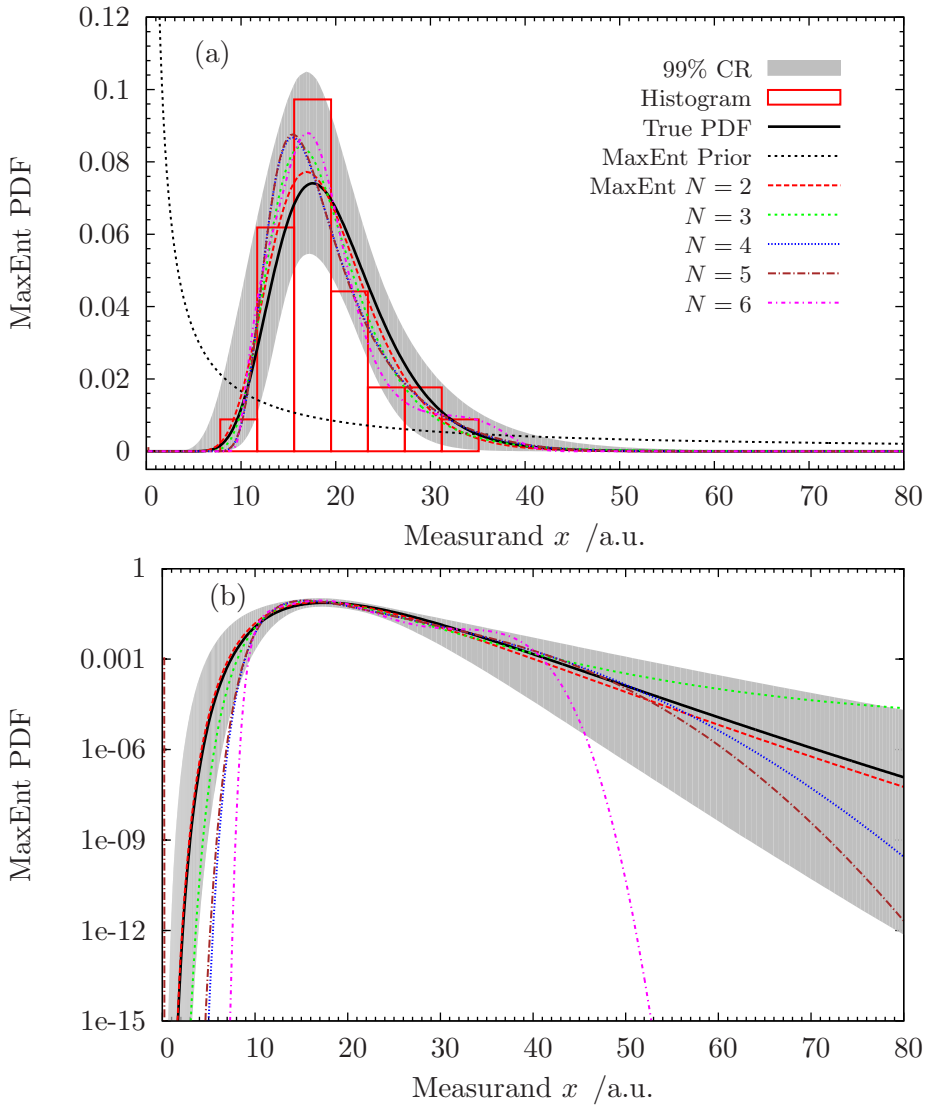

Figure 8: MaxEnt models from dataset of 30 samples, assuming a lognormal pdf: (a) Comparing the MaxEnt models with data, prior $m(x)$ and true pdf over a linear-scale. (b) Comparing MaxEnt models and true pdf over a logarithmic scale. The grey shaded area corresponds to the 99% credibility region for the $N = 2$ model (most probable model). All pdfs have been normalised to unit area.

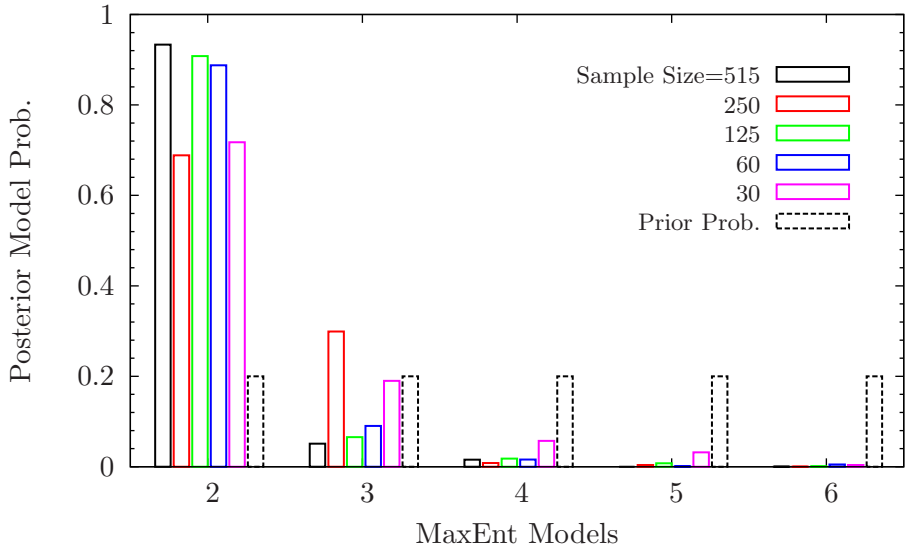

Figure 9: Posterior model probabilities for the log-normal test. MaxEnt models $M = 2$ to 6 and sample sizes: 30, 60, 125, 250 and 515. The true distribution in the transformed variable was set to a Gaussian pdf. In all cases the correct MaxEnt model, $N = 2$, was selected.

To assess the goodness of fit of each MaxEnt pdf, we use the information gain, denoted I, as a measure. I is the expected value, over the true pdf, of the logarithm of the ratio of the true pdf to the fitted pdf. That is, I (in bits) is given by,

$$I = \int f_{true}(x) \log_2 \left[f_{true}(x)/f_{maxent}(x) \right] \, \mathrm{d}x, \tag{25}$$

where as $f_{maxent}(x) \to f_{true}(x) \, \forall \, x$, then $I \to 0$.

The information gain is presented in Figure 10 for all the tests. For the Gaussian case, Figure 10(a), the fitted model with lowest I was $N = 2$ for all sample sizes, indicating that the MaxEnt pdf for $N = 2$ was closest to the true model for all samples considered. In this case, $N = 2$ was also selected as the best model structure for all sample sizes. As N increased, I also increased for each sample size, indicating that the noise was being modelled.

For the non-Gaussian case, Figure 10(b), for all sample sizes, I decreased from $N = 2$ to $N = 3$, indicating that there was enough data to model the third-order features of the true pdf. The information gain from $N = 3$ to $N = 4$ was positive for sample size 30, and negative for other sample sizes, decreasing as the sample size increased. This can be interpreted as the fourth-order features being subtle and requiring a large sample size to be truely represented; larger than 30, but with improvements still being made with up 515 samples. Adding a fifth moment made little difference to I, but adding the sixth moment increased I for all sample sizes and indicates that the noise, rather than the underlying true pdf, was starting to be modelled.

The lognormal case, Figure 10(c), was qualitatively similar to the Gaussian case, Figure 10(a), except the larger sample sizes did not model the noise until six moments were included.

5.5 Using other types of moments

The above approach can be applied to a wider class of MaxEnt model pdfs derived from different moments, other than polynomial moments, where an inference between competing MaxEnt models is shown to be necessary.

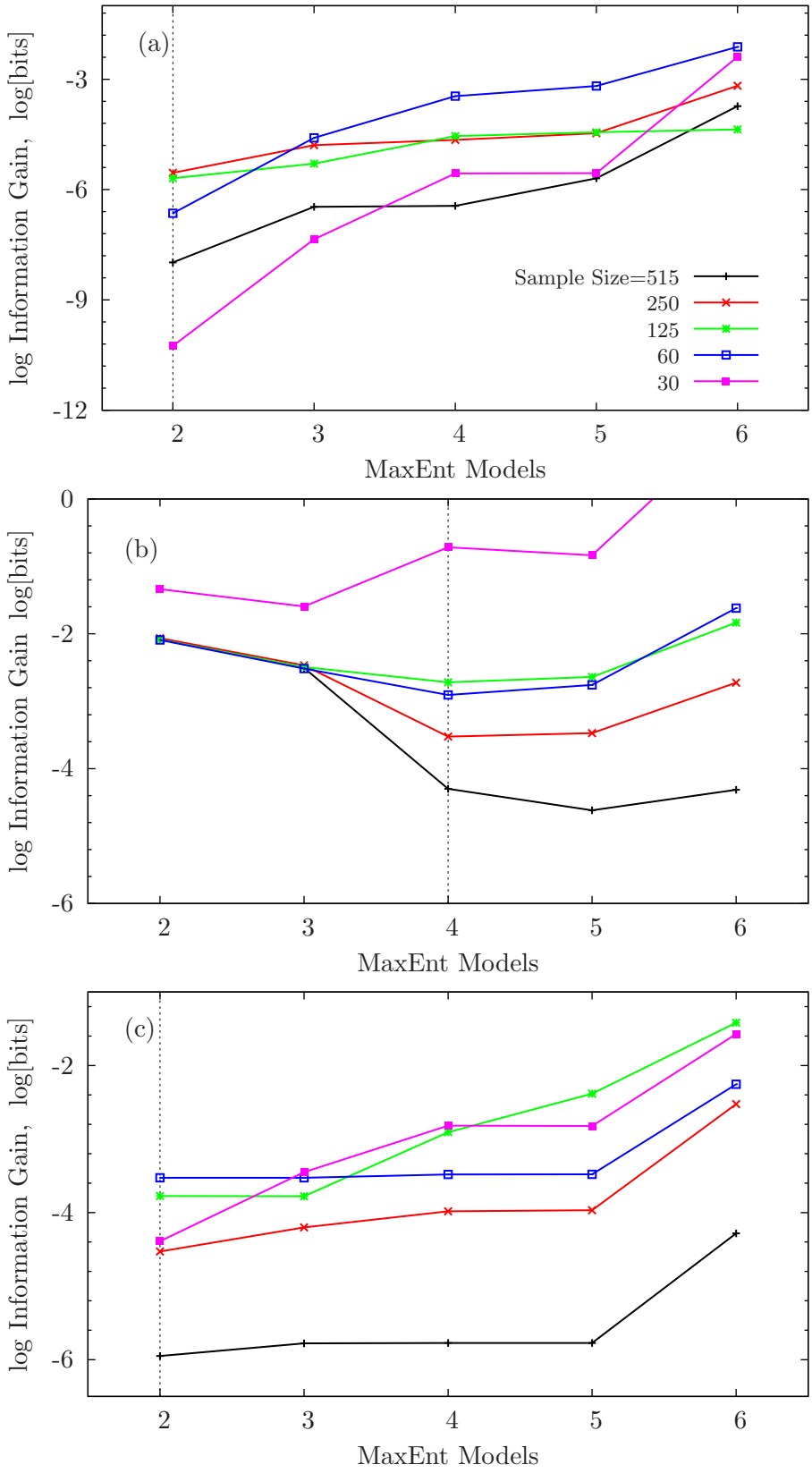

Figure 10: Information gain versus number of moment constraints, M, for sample sizes 30, 60, 125, 250 and 515. (a) Gaussian; (b) non-Gaussian; (c) lognormal cases.

For example, consider the case of two MaxEnt model pdfs defined by the same number of Lagrangian multipliers, but different constraints for the same data. The lognormal and Gamma pdfs models make different *a priori* assumptions about the same data (see Table 1 in [27]). At this point, the problem of which model best describes the data becomes one of inference and model selection. The equivalent free energy function (8a) for both models can be derived and used to estimate the Lagrangian multipliers. The prior for the Lagrangian multipliers will be defined by (18), while the Hessian (covariance) matrix can be determined from (16). Finally, the corresponding probabilistic evidence and posterior model probability can be calculated numerically using a MCMC (or equivalent) methods, as described in Sec. 3.4.

A key point that the current work demonstrates is that in order to assess competing MaxEnt models, Bayesian inference and model selection is necessary. Furthermore, both methods complement each other, where the MaxEnt method is used to determined the likelihood for the data that is necessary in the Bayesian model selection.

6 Conclusion

The objective of this work was to apply Bayesian model selection to determine the appropriate MaxEnt moments model conditional on the available data. The prior probability density function (pdf) for the Lagrangian multipliers, necessary to derive the probabilistic evidence, was derived by expanding the free-energy functional (16), where the location and covariance for the prior pdf were determined empirically as hyperparameters (18). A Markov Chain Monte Carlo method was used to numerically calculated the probabilistic evidence. Simulated datasets were used to test the combined MaxEnt moments/Bayesian approach for a range of sample sizes and distribution complexity. This involved generating random samples of data from pre-defined known distributions. Using the available data, the combined MaxEnt moments/Bayesian approach was assessed on how well it could reproduce the true distributions. The results demonstrated that the approach can successfully select the correct model from sample sizes $\gtrsim 30$ for simple models (i.e. Gaussian pdfs) and sample sizes of $\gtrsim 60$ for complex models (i.e. non-Gaussian pdfs), since the model structure is not known *a priori* we recommend a minimum sample size of 60 to distinguish between model structures of $N = 4$. In each case, the uncertainty region for the pdf of the most probable model structure was calculated and could be used to exclude competing models. Further exploration could generalise the limits of the approach, including higher order models or bimodal models, for example and its application in a model average procedure in order to address model uncertainty [14].

We also outlined how the combined MaxEnt moments/Bayesian approach could be generalised to assess competing MaxEnt models derived from moments, other than polynomial moments. A simple example was used to illustrated how Bayesian model selection would be necessary to determine the most probable MaxEnt model from a set of models that all had the same number of moments, but were derived from different assumptions about the same data. A key finding of our work is that the MaxEnt method for pdf assignment complements Bayesian model selection. Further work is required to assess the robustness of the approach and appropriately define a prior pdf for the number of moments necessary for calculating the posterior model probability. Finally, we have demonstrated that this approach is fully quantitative, and we believe it has many practical applications where a pdf must be estimated from limited data.

Acknowledgement

The authors would like to thank Dr. Paul Goggans, University of Mississippi, USA, for making us aware of Dr. Bretthorst's maximum entropy and Bayesian formulation on the 16th December 2013, at the MAXENT 2013 Conference, Canberra Australia, 15–12 December 2013. NA thanks Dr. Andy Rawlinson (IBM Melbourne) for useful discussions. This work was supported by DSTO 2013 Divisional Enabling Research Program Funding (NA) and ARC Discovery Project Grant No. DP110105290 (GJS & DBH).

A Overview of MCMC Procedure

A.1 Computational Details

For a given MaxEnt model, pre-analysis MCMC was carried out using algorithm 1 to tune the proposal pdf such that the acceptance ratio of the random walk resulted in "efficient" MCMC sampling [46–51]. The acceptance ratio was selected to maximise the efficiency of the MCMC sampling.

A total of 12 independent MCMC simulations were carried out for a given MaxEnt model. Each simulation consisted of 1.25×10^5 runs, where the first 20% was considered as the *burn-in* and rejected. Using the remaining 80% of the MCMC sample, the autocorrelation function of the probabilistic evidence was calculated and the autocorrelation length was determined. The post-burn-in MCMC sample (i.e. remaining 80%) was then re-sampled using the calculated autocorrelation length to remove the serial correlations in the MCMC sample. The re-sampled MCMC data from all 12 simulations were then augmented into a single MCMC dataset and used to determine the probabilistic evidence by applying the reverse importance sampling technique (see [32], Appendix E.4, pp. 220-222). Depending on the efficiency of the MCMC sampling ($\sim 2\%$ to 10%), the total number of data points in the single MCMC dataset ranged from 2.4×10^4 to 1.2×10^5. Once the probabilistic evidence was determined for all the models, the posterior model probability, (9), was evaluated, with normalisation according to (14).

Algorithm 1 Pseudocode for Metropolis Algorithm

Initialize $a = 0$ ▷ Initialize acceptance counter

Initialize $q(\boldsymbol{\lambda}|\,\boldsymbol{\lambda}_{(l)},\,\boldsymbol{\Sigma}_{pd})$ ▷ Initialize *proposal distribution*

Initialize Gaussian quadrature, $\{\tilde{\mathbf{x}},\,w\}$ ▷ Calculate Gaussian quadrature over $[\tilde{x}_{min},\,\tilde{x}_{max}]$, where $\tilde{\mathbf{x}} = \{\tilde{x}_k;\,k = 1,\,2,\,3,\ldots,\,K\}$ for K sample points.

Apply $\mathcal{T}:\,x \rightarrow \tilde{x}$ ▷ Define $\tilde{x}$-range, $\tilde{x} \in [\tilde{x}_{min},\,\tilde{x}_{max}]$ for K sample points.

Initialize $\boldsymbol{\lambda}_{(0)}$ ▷ Initialize starting value

Calculate $f_{(0)}(\tilde{\mathbf{x}}|\,\boldsymbol{\lambda}_{(0)},\,\mathcal{M}_N)$ ▷ Calculate the MaxEnt pdf using (6c)

Interpolate $f_{(0)}$ over $\tilde{x}$-range ▷ Interpolate (6c) over $\tilde{x} \in [\tilde{x}_{min},\,\tilde{x}_{max}]$ for for K sample points.

Calculate $p(\mathbf{D}|\boldsymbol{\lambda}_{(0)}\,\mathcal{M}_N,\,\mathcal{I})$ ▷ Calculate likelihood using (10).

Calculate $p(\boldsymbol{\lambda}_{(0)},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})$ ▷ Calculate the *joint-pdf* by using (18) and (10) in (11)

for $l \leftarrow 1, L_{max}$ **do**

  $\boldsymbol{\lambda}_{trial} \leftarrow q(\boldsymbol{\lambda}|\,\boldsymbol{\lambda}_{(l)},\,\boldsymbol{\Sigma}_{pd})$ ▷ Randomly draw $\boldsymbol{\lambda}_{trial}$

  Calculate $f_{trial}(\tilde{\mathbf{x}}|\,\boldsymbol{\lambda}_{trial},\,\mathcal{M}_N)$ ▷ Calculate new MaxEnt pdf using (6c)

  Apply $\mathcal{T}^{-1}:\,\tilde{x} \rightarrow x$ *and interpolate* f_{trial} *over* x*-range* ▷ Interpolate (6c) over $x \in [x_{min},\,x_{max}]$ for for K sample points.

  Calculate $p(\mathbf{D}|\boldsymbol{\lambda}_{trial}\,\mathcal{M}_N,\,\mathcal{I})$ ▷ Calculate likelihood using (10).

  Calculate $p(\boldsymbol{\lambda}_{trial},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})$ ▷ Calculate the *joint-pdf* by using (18) and (10) in (11)

  $w \leftarrow \frac{p(\boldsymbol{\lambda}_{trial},\mathbf{D}|\,\mathcal{M}_N,\mathcal{I})}{p(\boldsymbol{\lambda}_{(l)},\mathbf{D}|\,\mathcal{M}_N,\mathcal{I})}$ ▷ Evaluate the ratio, w, to compare $\boldsymbol{\lambda}_{trial}$ with $\boldsymbol{\lambda}_{(l)}$

  $r \leftarrow U[0,\,1]$ ▷ Randomly draw r from a uniform pdf [0, 1]

  if $r \leq w$ **then**

    $\boldsymbol{\lambda}_{(l+1)} = \boldsymbol{\lambda}_{trial};\,p(\boldsymbol{\lambda}_{(l+1)},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I}) = p(\boldsymbol{\lambda}_{trial},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})$ ▷ Only accept λ_{trial}, else reject.

    a=a+1

  else

    $\boldsymbol{\lambda}_{(l+1)} = \boldsymbol{\lambda}_{(l)};\,p(\boldsymbol{\lambda}_{(l+1)},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I}) = p(\boldsymbol{\lambda}_{(l)},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})$

  end if

  Store $p(\boldsymbol{\lambda}_{(l+1)},\,\mathbf{D}|\,\mathcal{M}_N,\,\mathcal{I})$ and $\boldsymbol{\lambda}_{(l+1)}$ ▷ Store values in an array

end for

$aratio = \frac{a}{L_{max}}$ ▷ Evaluate the acceptance ratio

References

[1] A. Azzalini. *Statistical inference based in the likelihood.* CRC Press, 1996.

[2] W. Q. Meeker and L. A. Escobar. *Statistical methods for reliability data.* Wiley, New York, 1998.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, New York, 2nd edition, 2003.

[4] D. J. C. MacKay. *Information theory, inference and learning algorithms.* Cambridge Uni. Press, Cambridge, 2003.

[5] D. S. Sivia and J. Skilling. *Data analysis: A Bayesian tutorial.* Oxford Uni. Press, Oxford, 2006.

[6] K. P. Murphy. *Machine learning: A probabilistic approach.* MIT Press, Cambridge, USA, 2012.

[7] L. G. Johnson. The choice of a class interval. *J. Amer. Stat. Assoc.*, 21(153):65–66, 1926.

[8] D. W. Scott. On optimal and data-based histograms. *Biometrika.*, 66(3):605–610, 1979.

[9] D. Freedmanfre and P Diaconis. On the histogram as a density estimator: l_2 theory. *Z.Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):4534760, 1981.

[10] E. Parzen. On estimation of a probability density function and mode. *The Annals Math. Stat.*, 33(3):1065–1076, 1962.

[11] B. W. Silverman. *Density estimation for statistics and data analysis*. CRC Press, 2nd edition, 1996.

[12] P. Hall, J. Racine, and Li Q. Cross-validation and the estimation of conditional probability density. *J. Amer. Stat. Assoc.*, 99(468):1015–1026, 2004.

[13] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984.

[14] J. A. Hoeting, D. Madingham, A. E. Raftery, and C. V. Volinsky. Bayesian model averaging: tutorial. *Statistical science*, 14(4):382–417, 1999.

[15] U. von Toussaint. Bayesian inference in physics. *Rev. Modern Phys.*, 7:943–999, 2011.

[16] Armstrong N. Hibbert, D. B. and J. H. Vine. Total CO_2 measurements in horses – where to draw the line. *Acced. Qual. Assur.*, 16:339–345, 2011.

[17] G. L. Bretthorst. The maximum entropy method of moments and Bayesian probability theory. In *AIP Conference proceedings*, volume 1553, 2013.

[18] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, 1957.

[19] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge Uni. Press, Cambridge, 2003.

[20] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and principle of minimum cross-entropy. *IEEE Trans. IT*, 26(1):26–37, 1980.

[21] R. W. Johnson and J. E. Shore. Comments on and correction to "Axiomatic derivation of the principle of maximum entropy and principle of minimum cross-entropy". *IEEE Trans. IT*, 26(6):942–943, 1983.

[22] P. Ishwar and P. Moulin. On the existence and characterisation of the maxent distribution under general moment inequality constrains. *IEEE Trans. Information Theory*, 51(19):3322–3333, 2005.

[23] S. Pressé, G. Kingshuk, J. Lee, and K. A. Dill. Nonadditive entropies yield probability distributions with biases not warranted by the data. *Phys. Rev. Lett.*, 11:180604–180609, 2013.

[24] S. Pressé, G. Kingshuk, J. Lee, and K. A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.*, 85:1115–1141, 2013.

[25] W. Wu. Calculation of maximum entropy densities with application to income data. *J. Econometrics*, 115:347–354, 2003.

[26] M. Rockinger and E. Jondeau. Entropy densities with application to autoregressive conditional skewness ad kurtosis. *J. Econometrics*, 106:119–142, 2002.

[27] S. Y. Park and A. K. Bera. Maximum entropy autoregressive conditional heteroskedasticity model. *J. Econometrics*, 150:219–230, 2009.

[28] R. V. Abramov. The multidimensional maximum entropy moment problem: A review of numerical methods. *Comm. Math. Sc.*, 8:377–392, 2010.

[29] S. Ciulli, M. Mounsif, N. Gorman, and T.D. Spearman. On the application of maximum entropy to the moments problem. *J. Math. Phys.*, 32(7):1717–1719, 1991.

[30] A. Zellner and J. Tobias. Further results on Bayesian method of moments analysis of the multiple regression model. *Int. Econ. Rev.*, 42(1):121–139, 2001.

[31] R. E. Kass and A. E. Raftery. Bayes factor. *J. Amer. Stat. Assoc.*, 90(430):773–795, 1995.

[32] J. J. K. O'Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian methods applied to signal processing*. Springer, New York, 1996.

[33] G. D'Agostini. Bayesian inference in processing experimental data: principles and basic applications. *Rep. Prog. Phys.*, 66:1383–1419, 2003.

[34] V. Dose. Bayesian inference in physics: case studies. *Rep. Prog. Phys.*, 66:1421–1461, 2003.

[35] P. C. Gregory. *Bayesian logical data analysis for the physical sciences*. Cambridge Uni. Press, Cambridge, 2005.

[36] J. Skilling. Nested sampling for General Bayesian Computations. *Bayesian Analysis*, 1(4):833–860, 2006.

[37] N. Friel and J. Wyse. Estimating the model evidence: A review. arXiv: 1111.1957, 26 November 2011.

[38] M. D. Weiburg. Computing the Bayes Factor from Markov Chain Monte Carlo simulations of the Posterior distribution. *Bayesian Analysis*, 7(3):737–770, 2012.

[39] E. T. Jaynes. Prior probabilities. *IEEE Trans. SSC*, 4(3):227–240, 1968.

[40] N. Wu. *The Maximum Entropy Method*. Springer-Verlag, Berlin, 1997.

[41] W.R Gilks, S. Richardon, and D. J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

[42] M. A. Newton and A. E. Raftery. Approximate Bayesian inference with weighted likelihood bootstrap. *J. R. Statist. Soc. B*, 56(1):3–48, 1994.

[43] A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *J. R. Statist. Soc. B*, 56(3):501–514, 1994.

[44] J. Skilling. Nested sampling for Bayesian Computations. In *ISBA 8th World Meeting on Bayesian Statistics*, pages 1–25, Alicante, Spain, 1–6 June 2006.

[45] M. Daghofer, M. Konegger, H. G. Evetz, and W. von der Linden. Perfect tempering. arXiv: 0512167, 22 May 2006.

[46] P. Neal and G. O. Roberts. Optimal scaling of random walk Metropolis-Hastings algorithms with non-Gaussian proposal. *Methodol. Comput. Appl. Probab.*, 13:583–601, 2011.

[47] G. O. Roberts and J. S. Rosenthal. Optimal scaling of Metropolis-Hastings algorithms: is 0.234 as robust as is believed. 2007.

[48] S. F. Jarnet and G. O. Roberts. Convergence of heavy-tailed Monte Carlo Markov Chain algorithms. *Scan. J. Stat.*, 34:781–815, 2007.

[49] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sc.*, 16(4):351–367, 2001.

[50] G. O. Roberts, G. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals App. Probab.*, 7(1):110–120, 1997.

[51] K. M. Hanson and G. S. Cunningham. Posterior sampling with improved efficiency. *SPIE*, 3338:371–382, 1998.