OCTOBER 07 2024

Maximum entropy temperature selection via the equipartition theorem FREE

Jacob R. Nuttall; Tracianne B. Neilsen ; Mark K. Transtrum

Proc. Mtgs. Acoust. 52, 070004 (2023)
https://doi.org/10.1121/2.0001959

Volume 52 http://acousticalsociety.org/

Acoustics 2023 Sydney

185th Meeting of the Acoustical Society of America

Sydney, Australia

4-8 December 2023

Underwater Acoustics: Paper 1aUW5

Maximum entropy temperature selection via the equipartition theorem

Jacob R. Nuttall, **Tracianne B. Neilsen and Mark K. Transtrum**

Department of Physics and Astronomy, Brigham Young University, Provo, UT, 84602; jacob_nuttall@outlook.com; tbn@byu.edu; mark.transtrum@byu.edu

Maximum entropy is an approach for obtaining posterior probability distributions of modeling parameters. This approach, based on a cost function that quantifies the data-model mismatch, relies on an estimate of an appropriate temperature. Selection of this ”statistical temperature” is related to estimating the noise covariance. A method for selecting the ”statistical temperature” is derived from analogies with statistical mechanics, including the equipartition theorem. Using the equipartition-theorem estimate, the statistical temperature can be obtained for a single data sample instead of via the ensemble approach used previously. Examples of how the choice of temperature impacts the posterior distributions are shown using a toy model. The examples demonstrate the impact of the choice of the temperature on the resulting posterior probability distributions and the advantages of using the equipartition-theorem approach for selecting the temperature.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0001959

1. INTRODUCTION

The approach of maximizing entropy for inference problems in an information theoretic modeling setting was first introduced by E.T. Jaynes in his 1957 papers.[1,2] The objective of Jaynes' work was to propose a rigorous method for choosing unbiased probability distributions $P(x)$ on a random variable x over a hypothesis space Ω_x. Jaynes' method uses the maximum entropy (MaxEnt) principle, which aims to make as few assumptions as possible to objectively derive probability distributions.[3] MaxEnt maximizes the information entropy,[4] $S = -\Sigma_{x\in\Omega_x} P(x)\log P(x)$, subject to two constraints: (1) the distribution $P(x)$ is normalized to unity over Ω_x, making it a probability density function: $\Sigma_{x\in\Omega_x} P(x) = 1$; and (2) for a function $f(x)$, the expectation value $\langle f\rangle = \Sigma_{x\in\Omega_x} P(x)f(x)$ is known. These two constraints, as well as the choice of hypothesis space, represent the minimal assumptions that can be made about a probability distribution.

Maximizing information entropy in this manner results in probability distributions analogous to those obtained in statistical mechanics for the energies of a system of particles held at a fixed temperature. Through this analogy, insights from thermodynamics can be extended to inference problems. It has proven especially useful in Bayesian inference, where the objective is to sample from the posterior probability distribution (PPD). In this context, MaxEnt is most commonly used to select uninformative priors in Bayesian inference;[5–7] however, it can similarly motivate the likelihood function.[8,9]

The purpose of this paper is to explain the analogies with statistical mechanics, especially the equipartition theorem, for inference problems in underwater acoustics. We argue that the noise variance is analogous to temperature in a thermodynamic system and use this insight to estimate the noise variance for a single data sample. An example is provided using an analytical model to highlight how the PPDs depend on the selection of the noise variance, i.e., temperature, especially when only a single measurement is available.

2. METHODS

Inferring the parameters of a model from measured data requires a cost function (sometimes called the loss or error function) that quantifies the data-model mismatch. The best fit parameters are those that minimize the cost function. In this paper, model predictions are represented by the variable $\mathcal{F}$ and are functions of the N modeling parameters in $\boldsymbol{\theta}$. The modeled values $\mathcal{F}(\boldsymbol{\theta})$ are compared to measured values (observations) $\mathcal{D}$ via a cost function. The cost function is denoted by ϵ to emphasize that it is analogous to energy in statistical mechanics. For this example, the cost function is the sum of squared error (SSE), which is often denoted by chi-squared χ^2 in the statistics literature:

$$\epsilon(\boldsymbol{\theta},\mathcal{D}) = \chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{M} \mathcal{R}_i(\boldsymbol{\theta},\mathcal{D})^2, \tag{1}$$

where the residuals $\mathcal{R}_i$ are defined as

$$\mathcal{R}_i(\boldsymbol{\theta},\mathcal{D}) = \mathcal{F}_i(\boldsymbol{\theta}) - \mathcal{D}_i. \tag{2}$$

The subscript i corresponds to the ith observation of $\mathcal{D}$, and M is the total number of observations.

The inference process considers the measured data as constant and samples the model input parameters $\boldsymbol{\theta}$ over predefined bounds, for example by Monte Carlo (MC) sampling. Here, each parameter θ_j is assumed to be a uniform random variable distributed as $\theta_j \sim \mathcal{U}(\theta_{j,a},\theta_{j,b})$, where $\theta_{j,a}$ and $\theta_{j,b}$ are the lower and upper bounds for that parameter, respectively. This assumption yields an uninformative prior because every value within the selected bounds is equally likely and corresponds to the MaxEnt prior for these constraints.

Residuals $\mathcal{R}_i$ exist for different reasons. First, measurements $\mathcal{D}$ contain noise. Noise in this context is not referring to ambient acoustical sounds but rather to measurement noise that may be assumed to be

independent, zero-mean Gaussian distributed random variables. Second, biases in measurements may occur. Third, the model producing $\mathcal{F}_i(\boldsymbol{\theta})$ may not capture all physical processes, which leads to model error. Throughout the example provided in this paper, the term noise refers to the first category of independent and identically distributed (IID) Gaussian random noise that occurs due to the measurement process.

A. LIKELIHOOD FUNCTION

The likelihood function for a model $\mathcal{F}$ is obtained by first assuming each of the M residuals $\{\mathcal{R}_i(\boldsymbol{\theta}, \mathcal{D})\}$ are IID Gaussian random variables with variance σ^2 and zero mean. With this assumption, the joint PDF for all of the residuals is given by

$$p(\mathcal{R}_1, ..., \mathcal{R}_M|\boldsymbol{\theta}, \mathcal{D}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^M \exp\left\{-\frac{1}{2\sigma^2}\epsilon(\boldsymbol{\theta}, \mathcal{D})\right\}. \tag{3}$$

The statistical formulation of the joint distribution of residuals in Eq. 3 depends on $\boldsymbol{\theta}$ and observed data $\mathcal{D}$. Because $\mathcal{D}$ is fixed and the input parameters $\boldsymbol{\theta}$ are varied in an inference problem, this PDF indicates the likelihood that $\mathcal{D}$ was generated by parameteris $\boldsymbol{\theta}$. The likelihood function from Eq. 3 is

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\left\{-\frac{1}{2\sigma^2}\epsilon(\boldsymbol{\theta}, \mathcal{D})\right\}. \tag{4}$$

The form of Eq. 3 implies that $\epsilon(\boldsymbol{\theta}, \mathcal{D})$ may also be treated as a random variable. Indeed, ϵ is distributed according to a χ-squared distribution, which is the reason it is often simply called χ-squared in the statistics literature.

The likelihood is not fully specified without estimating a value for σ^2. If measurements are repeated, noise variance can be empirically estimated directly from the observations; this apporach, however, is not possible when there is only a single measurement to work with. The next section begins to approach the problem of estimating σ^2 by constructing an analogy with statistical mechanics using the maximum entropy (MaxEnt) principle.

B. PRINCIPLE OF MAXIMUM ENTROPY

i. Maximizing Information Entropy

For a discrete hypothesis or parameter space $\Omega_x = (x_1, ..., x_n)$, with an unknown associated probability distribution $p = (p_1, ..., p_n)$, the MaxEnt principle selects the distribution p which maximizes the Shannon information entropy given by

$$S(p) = -\sum_{i=1}^{n} p_i \log p_i, \tag{5}$$

subject to two constraints.[1,2] The first constraint is that p is normalized: $\sum_{i=1}^{n} p_i = 1$. The second constraint is that for some function $f(x)$, the expectation value $\langle f \rangle = \sum_{i=1}^{n} p_i f(x_i)$ is known. Additional constraints may be added for if the expectation value of other functions is also known. The constraints involving expectation values constitute known information, which is generally insufficient to uniquely constrain each p_i. The full probability distribution can be determined, however, by maximizing the entropy subject to these constraints. In practice, the optimization procedure is done using the method of Lagrange multipliers.[10] The resulting p_i is given by

$$p_i = \frac{1}{Z} e^{-\beta f(x_i)}, \tag{6}$$

where Z is the Lagrange multiplier associated with the normalization constraint. Z is analogous to the partition function in statistical mechanics and is given by

$$Z = \sum_{i=1}^{n} e^{-\beta f(x_i)}. \tag{7}$$

The second Lagrange multiplier, β, is associated with the constraint $\langle f \rangle$.

In most inference problems, the parameter space Ω_x of interest is continuous, rather than discrete. In this case, the two constraints generalize as $\int_{\Omega_x} p(x)dx = 1$ and $\langle f \rangle = \int_{\Omega_x} f(x)p(x)dx$. The information entropy for a continuous variable x is defined relative to a reference measure $q(x)$:

$$S(p) = -\left\langle \log \frac{p(x)}{q(x)} \right\rangle = -\int_{\Omega_x} p(x) \log \frac{p(x)}{q(x)} dx. \tag{8}$$

When $q(x)$ is also a probability distribution (i.e., a normalized measure), the quantity $S(p)$ is also known as the Kullback-Leibler divergence or relative entropy.[11] If the reference measure $q(x)$ is a uniform distribution, then the probability distribution that maximizes entropy is given by[12]

$$p(x) = \frac{1}{Z} e^{-\beta f(x)}, \tag{9}$$

where, as before, Z and β are the Lagrange multipliers associated with the normalization and $\langle f \rangle$ respectively. The partition function Z is given by

$$Z = \int_{\Omega_x} e^{-\beta f(x)} dx. \tag{10}$$

Jaynes original motivation was for the MaxEnt principle for selecting minimally informative priors consistent with available information.[5,6] A recent tutorial by Xiang[7] has provided an excellent review of the relationship between MaxEnt and establishing priors for Bayesian inferences. In addition to prior selection, MaxEnt may be used as a method for obtaining PPDs.[8,9] To obtain PPDs using MaxEnt, however, the noise covariance σ^2 must be estimated.

Previous work in geoacoustic inversions have used an ensemble approach to estimating σ^2. This ensemble approach was used by Stotts and Koch in[12] as well as Knobles et al. in,[13,14] in which multiple ship noise data samples were used to estimate σ^2 for geoacoustic inversion. In this approach, the likelihood function represents the change in information due to the addition of data $\mathcal{D}$, quantified by a known expectation value for a function f, updates beliefs about parameters $\boldsymbol{\theta}$.

For a continuous variable, the reference measure $q(\boldsymbol{\theta})$ is the Bayesian prior. That is, the MaxEnt likelihood maximizes relative entropy subject to the aforementioned constraints over a parameter space with prior distribution $q(\boldsymbol{\theta})$. For uniform priors, the likelihood obtained from the MaxEnt principle has a nice relationship to the likelihood obtained in the statistical formulation. In particular, extending the analogy with statistical mechanics gives an estimate for the noise variance, σ^2.

ii. Multi-Particle Partition Function

In statistical mechanics, the partition function is a quantity from which all of the relevant thermodynamics can be derived, such as the internal energy, pressure, heat capacity, and chemical potential. For a system of many particles (such as an ideal gas) with energy levels E_i in equilibrium with a thermal reservoir at temperature T, the partition function $Z_{\text{particles}}$ is given by a sum of the Boltzmann factors corresponding to each available energy level:

$$Z_{\text{particles}} = \sum_{i=1}^{n} e^{-E_i/k_B T}, \tag{11}$$

where k_B is Boltzmann's constant. This partition function is known as the *canonical partition function* , and the system is represented by a statistical ensemble called the *canonical ensemble*.[15,16]

iii. Statistical Partition Function

Parameter values $\boldsymbol{\theta}$ of a model $\mathcal{F}$ associated with lower cost values $\epsilon(\boldsymbol{\theta}, \mathcal{D})$ are in better agreement with the measured data $\mathcal{D}$ and, therefore, inferred to be more likely. The minimum, best fit value of the cost is denoted by ϵ_0. Statistical inference characterizes parameter values such that their associated cost, $\epsilon(\boldsymbol{\theta})$ is not too much larger than ϵ_0. How much larger than ϵ_0 while still being statistically significant is controlled by the choice of variance σ^2.

We apply the MaxEnt principle with $f = \epsilon$ in Eq. 9. In the thermodynamic analogy, this choice corresponds to a system in thermal equilibrium with a reservoir. The Lagrange multiplier introduced by this choice determines the average value of the cost, just as the temperature determines the average energy. In a physical system, the temperature also sets the scale of energy *fluctuations*, $\sim k_B T$.

In similar fashion, applying MaxEnt sets the scale for the statistical fluctuations in the cost and determine σ^2. In this application, k_B is assumed to be one, and T is thought of as a *statistical temperature* related to the information. This T is commonly referred to as temperature.[17] Having made this choice, it remains to select a value for $\langle \epsilon \rangle$, i.e., what T should be used? This question will be addressed in Sec. 2.3.

Assuming that the parameter space $\Omega_{\boldsymbol{\theta}}$ is continuous and that the reference measure $q(\boldsymbol{\theta})$ is uniform, the partition function in Eq 10, for this case, is given by

$$Z_{\text{stat}} = \int_{\Omega_{\boldsymbol{\theta}}} e^{-\beta \epsilon(\boldsymbol{\theta}, \mathcal{D})} d\boldsymbol{\theta}, \tag{12}$$

which is referred to as the *statistical partition function* for the model $\mathcal{F}$.

The interpretation of the quantities in the canonical partition function may be drawn upon to assign meaning to both the cost function $\epsilon(\boldsymbol{\theta}, \mathcal{D})$ as well as the Lagrange multiplier β. By comparing the statistical and canonical partition functions, the quantity ϵ corresponds to the energy as in Eq. 11 and, thus, is called the *statistical energy* of the model $\mathcal{F}$ for parameterization $\boldsymbol{\theta}$. Assuming $k_B = 1$, the parameter $\beta = 1/T$ in the statistical partition function may be recognized as analogous to the inverse temperature. and is referred to as the *statistical temperature* of the model $\mathcal{F}$.

iv. Derivation of the Likelihood

With the above results, and treating the distribution obtained from the MaxEnt principle as a posterior distribution $p(x) \rightarrow p(\boldsymbol{\theta}|\mathcal{D})$ for parameters $\boldsymbol{\theta}$ and observed data $\mathcal{D}$, the probability distribution in Eq. 9 becomes

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z_{\text{stat}}} e^{-\beta \epsilon(\boldsymbol{\theta}|\mathcal{D})}, \tag{13}$$

where the statistical partition function Z_{stat} acts as a normalization constant.
A uniform reference measure $q(x)$ is assumed in the derivation of the partition function Z_{stat} in Eq 13, so the priors $q(x) \rightarrow p(\boldsymbol{\theta})$ are also uniform over specified bounds $(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$. With these changes, the likelihood function may be written in terms of the statistical temperature $T = \beta^{-1}$ with

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\left\{ -\frac{1}{T} \epsilon(\boldsymbol{\theta}, \mathcal{D}) \right\}. \tag{14}$$

Comparison of Eq. 14 with the likelihood obtained from the statistical formulation as in Eq. 4 indicates that the statistical temperature is proportional to the noise covariance of the data σ^2, which is an estimate of the total errors in the model and data:

$$T = 2\sigma^2. \tag{15}$$

Because $\beta = 1/T = 1/2\sigma^2$ is proportional to the information content of the data, we see that high temperature implies low information. This matches our physical intuition in which high physical temperature gives little information about the microstate of the system, while low temperatures are more likely to be found in the ground state. We now use this physical intuition to determine a natural scale for T in the following section.

C. TEMPERATURE ESTIMATION

The ultimate goal of using the MaxEnt principle in this paper is to estimate for the noise covariance σ^2 using an analogy with statistical mechanics. In this section we use the equipartition theorem to motivate a natural choice of statistical temperature .

The equipartition theorem states that, on average, the internal energy U of a system is distributed evenly among the available degrees of freedom, with each degree of freedom having an average energy of $\frac{1}{2}k_BT$. In the context of a thermodynamic system at constant temperature, the internal energy fluctuates due to thermal exchange with a heat bath. Using the PPD obtained in Eq. 13, the average statistical energy is

$$U = \langle \epsilon \rangle = \int_{\Omega_{\boldsymbol{\theta}}} \epsilon(\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}. \tag{16}$$

While this quantity is unknown, the minimum cost function, $\epsilon_0 = \min_{\boldsymbol{\theta}} \epsilon(\boldsymbol{\theta}, \mathcal{D})$, is a lower bound and sets a natural scale for energy fluctuations.

Application of the equipartition theorem results in

$$\epsilon_0 \leq U = \sum_{i=1}^{N} \frac{T}{2} = \frac{N}{2} T. \tag{17}$$

Therefore, we advocate (as recommended in Frederiksen *et al.*[17]) setting the statistical temperature according to the minimum cost as

$$T \approx \frac{2}{N} \min \epsilon_0. \tag{18}$$

This method of applying the equipartition theorem to obtain an estimate of the statistical temperature has also been suggested in Refs. 18, 19. Temperature selection for MaxEnt is similar to Gibbs' sampling at different temperatures[20] and other work associated with choosing a data covariance matrix.[21,22]

Characterization of posterior distributions requires drawing a large number of samples from the PPD. A basic approach for selecting the samples is Monte Carlo (MC) sampling, which draws parameters $\boldsymbol{\theta}$ from the specified priors, evaluates the model at those parameter values, and then calculates $\epsilon(\boldsymbol{\theta}, \mathcal{D})$. For a large number of samples N_s, both the costs and associated parameter values are stored in memory. Because uniform priors are used, the joint posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ is proportional to the likelihood in Eq. 14.

i. Marginalizing the Joint Posterior

To numerically evaluate the marginal PPDs, the parameter space for the ith parameter, Ω_{θ_i}, is divided into B uniformly spaced bins. Then, the marginal probability density is calculated by averaging the likelihood function for the samples with θ_i that fall within each bin, using the Savitzky-Golay filter to smooth artifacts.[23]

The marginal PPDs are calculated as

$$\tilde{p}(\theta_{i,k}|\mathcal{D}) = \left(\sum_{|\theta_{i_j} - \theta_{i,k}| \leq \frac{1}{2}\Delta\theta_i} 1 \right)^{-1} \left(\sum_{|\theta_{i_j} - \theta_{i,k}| \leq \frac{1}{2}\Delta\theta_i} e^{-(1/T)\epsilon(\theta_i, \mathcal{D})} \right), \tag{19}$$

where $\theta_{i,k}$ corresponds to the kth bin center in the parameter space for the ith parameter, $\Delta\theta_i$ is the bin width of the B bins, and θ_{i_j} is the jth value drawn for the ith parameter. The Savitsky-Golay filter is then applied to this result with a window size of 10% the number of bins B and normalized, so that

$$p(\theta_{i,k}|\mathcal{D}) = \frac{1}{Z}\mathrm{SG}_{3,B/10}\left(\tilde{p}(\boldsymbol{\theta}_i|\mathcal{D})\right)_k . \tag{20}$$

In this equation, $\mathrm{SG}_{3,B/10}(\cdot)$ represents the Savitzky-Golay filter[23] applied with 3rd-order polynomials and a window size with $B/10$ points. $\boldsymbol{\theta}_i$ is a vector of the parameter values at the B bin centers for the ith parameter, and Z is the normalization constant so that the probability in Eq. 20 integrates to one. After applying the filter, particularly sharp portions of the PPD may have negative values. If this occurs, these negative values are zeroed out. Z is obtained by integrating the marginal PPD over Ω_{θ_i} using the composite trapezoidal rule as implemented in the `trapz` function in the `numpy` python library.

While the Savitsky-Golay filter yields reasonable marginal PPDs in this paper, a more common approach is to use a kernel-smoothed density estimator (KDE). A KDE should be used in future work as a more statistically robust method for smoothing estimates of PPDs.

D. SUMMARY STATISTICS

The marginal PPDs can be analyzed to obtain summary statistics such as the peak (mode), mean, and median parameter values for θ_i can be computed, as well as the 95% credibility interval (CI). The optimal parameters, or the parameters which minimize the cost function, are also identified.

The peak, mean, and optimal parameter values for the ith parameter are calculated as

$$\begin{aligned}
\theta_{i,\mathrm{peak}} &= \mathrm{argmax}_{\theta_i}\, p(\theta_i|\mathcal{D}), \\
\theta_{i,\mathrm{mean}} &= \int_{\theta_{i,a}}^{\theta_{i,b}} \theta_i p(\theta_i|\mathcal{D}) d\theta_i, \\
\theta_{i,\mathrm{optimal}} &= \mathrm{argmin}_{\theta_i}\, \epsilon(\theta_i, \mathcal{D}),
\end{aligned}$$

where $\theta_{i,a}$ and $\theta_{i,b}$ correspond to the lower and upper bound for θ_i defined in the priors. Because uniform priors are assumed, the optimal parameter values correspond to both the maximum likelihood estimate (MLE) and maximum *a posteriori* estimate (MAP). However, the MAP estimate is for the joint distribution of all of the parameters $\boldsymbol{\theta}$, so the MAP estimate for θ_i may not correspond to the maximum (peak) of its marginal distribution.

The median and 95% CI are calculated from the cumulative distribution function (CDF), which represents how much of the probability falls *below* a particular value for θ_i, say x. The CDF is given by

$$\mathrm{CDF}(x) = P(\theta_i \leq x|\mathcal{D}) = \int_{\theta_{i,a}}^{x} P(\theta_i|\mathcal{D}) d\theta_i. \tag{21}$$

From the CDF, the median is calculated as the point satisfying

$$\mathrm{CDF}(\theta_{i,\mathrm{median}}) = 0.5. \tag{22}$$

Similarly, the 95% CI, represented as $(\theta_{i,\mathrm{lower}}, \theta_{i,\mathrm{upper}})$ is obtained from the points which satisfy

$$\begin{aligned}
\mathrm{CDF}(\theta_{i,\mathrm{lower}}) &= 0.025, \\
\mathrm{CDF}(\theta_{i,\mathrm{upper}}) &= 0.975.
\end{aligned}$$

The CI is used because it may be considered as a better notion of the variability of a distribution, as it can be used to measure how wide or narrow is a probability distribution and more easily accounts for multimodal distributions than the standard deviation.

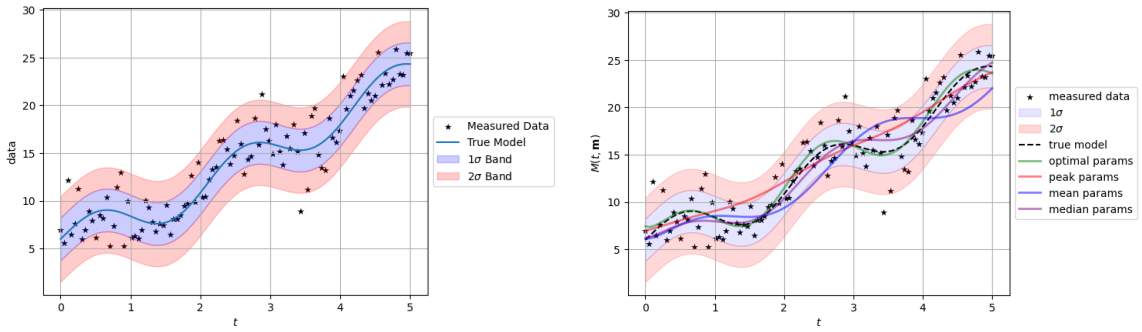

Figure 1: (Left) Visualization of the toy model defined in Eq. 23 used to illustrate the MaxEnt inference process: the underlying model $\mathcal{F}(t;\boldsymbol{\theta}_{\text{true}})$ ***(blue lines) with*** $\theta_{\text{true}} = [2, 3, 3, 4, 1]$***, the* measured *measured*** $\mathcal{D}$ ***with*** $\sigma^2 = 5$ ***(black stars), along with*** 1σ ***and*** 2σ ***uncertainty bands. (Right) Results of the MaxEnt in Sec. 3.2: Comparison of*** $\mathcal{F}(t;\boldsymbol{\theta})$ ***evaluated with*** θ ***at the optimal, peak, mean, and median values from the PPDs shown in Fig. 2.***

3. EXAMPLE

A. TOY MODEL

The maximum entropy (MaxEnt) approach is illustrated using a five-parameter toy model $\mathcal{F}$ built as a linear combination of a sine wave, a linear term, and a decay term. The modeling parameters $\boldsymbol{\theta} = [A, B, C, D, E]$ define the contributions of each of these terms:

$$\mathcal{F}(t;\boldsymbol{\theta}) = A\sin(Bt) + C + Dt + 3e^{-Et}, \tag{23}$$

where t is the independent variable.

Noisy *measured* data is simulated as

$$\mathcal{D}(t) = \mathcal{F}(t;\boldsymbol{\theta}_{\text{true}}) + \xi(t). \tag{24}$$

$\boldsymbol{\theta}_{\text{true}}$ contains the parameters used in generating $\mathcal{F}(t;\boldsymbol{\theta}_{\text{true}})$; for the example in this paper, $\boldsymbol{\theta}_{\text{true}}$ contains A = 2, B = 3, C = 3, D = 4, and E = 1. $\mathcal{F}(t;\boldsymbol{\theta}_{\text{true}})$ is evaluated at 100 evenly-spaced time points on the interval t = [0, 5], yielding the solid blue line in Fig. 1(a). $\mathcal{D}$ is obtained by evaluating the model at the true parameter values and adding Gaussian-distributed noise $\xi(t) \sim \mathcal{N}(0, \sigma^2)$. For this example, variance $\sigma^2 = 5$ is used, which corresponds to $T_{\text{true}} = 10$ (Eq. 15). The resulting *measured* data $\mathcal{D}$ are shown as black stars in Fig. 1(a). The colored areas indicate regimes of $\pm 1\sigma$ (purple) and $\pm 2\sigma$ (red).

$\mathcal{D}$ exhibits a strong linear trend due to the slope D, as well as the frequency B of the oscillations. The amplitude of the oscillations A, however, is swamped out by the added noise. Additionally, the effects of constant C and decay term $3e^{-Et}$ are not easily identifiable in $\mathcal{D}$. Therefore, we expect the MaxEnt process to yield PPDs that are informative for B and D and uninformative for A, C, and E.

B. MARGINAL PPDS

MaxEnt is applied to the toy model, using the analogy to the equipartition theorem, to obtain marginal PPDs for the five modeling parameters. MC sampling of $\mathcal{F}(t;\boldsymbol{\theta})$ (in Eq. 23) is performed using 100,000 samples. The samples of $\boldsymbol{\theta}$) are randomly selected from uniform bounds $\theta_i = [0, 5]$. The cost function is calculated for each sample (Eq. 1), and Eq. 18 yields $T = 1.686$. This estimate is used to calculate the likelihood function and the normalized marginal PPDs (Eq.19).

Marginal PPDs are computed using the method described in Sec. 2.3.1 and displayed for the five parameters in Fig. fig:posteriorvisual. The PPDs show that the parameter D is the most identifiable parameter, with

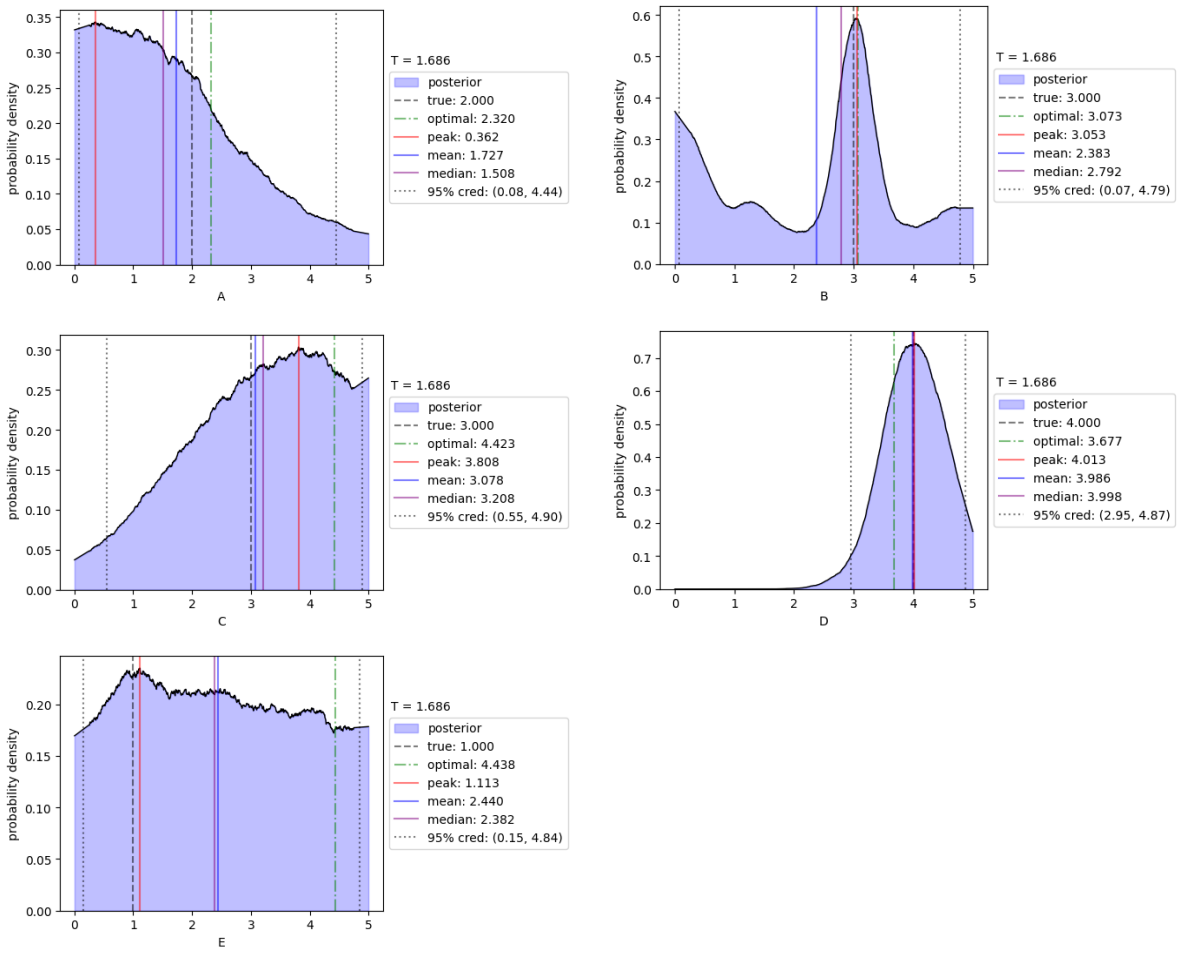

Figure 2: The marginal PPDs along with lines indicating the optimal, peak, mean, and median parameter values of $\theta = [A, B, C, D, E]$, as well as 95% CI. The statistical temperature of the model $T = 1.686$ in estimated using Eq. 18.

a narrow probability distribution and small 95%. The resulting optimal, peak, mean, and median values are close together and match $D_{\text{true}} = 4$. The PPD for parameter B is somewhat informative, due to the narrow highest peak. In this case, the optimal and peak values are close to $B_{\text{true}} = 3$. The 95% CI, however, is wide due to the multi-modality nature of the PPD indicating underlying uncertain in the estimated values.

The PPD for C is not multi-modal; the mean, and median parameter values are close to $C_{\text{true}} = 3$, while the optimal and peak values are higher. This analysis is only possible because the true values are known. In general, the large 95% CI would indicate only that C has a greater than 0.55 and is likely between 3 and 4. The posterior distribution for the parameter A shows a similar situation.

Finally, the PPD for E is essentially flat, with a wide 95% CI as well as having the optimal, peak, mean, and median parameters exhibiting a large spread. These features means that the data contains no information about the parameter E in the model, and any estimate of E inferred has a large uncertainty.

C. TEMPERATURE DEPENDENCE

The choice of the statistical temperature T has a significant impact on the resulting PPDs. Too low of a temperature results in sharply peaked probability distributions which may imply more information exists

from the data for a model then may be actually the case. On the other hand, too high of a temperature may imply that the data is uninformative. The statistical perspective, as in Sec. 2, provides insight into an appropriate choice for T and how it relates to the noise variance.

The estimate of T should match the informativity of the data. Since the variance of the data is unknown in practice, especially with only one data sample, the temperature must be estimated. This toy model provides evidence that the approach of estimating T using the equipartition theorem analogy (Eq. 18) obtains posterior distributions that capture features of the model at the T_{true}.

The impact of T on PPDs for the five parameters in the toy model are shown in Fig. 3. The first row demonstrates what happens in the low temperature regime by using $T = 0.1 << T_{\text{true}}$: narrow, sharply peaked distributions appear with low variance. This implies greater information exists for the parameters than evidenced by the noisy data. The second row uses $T = 1.68$, estimated from Eq. 18. In this row, the probability distributions are a reasonable representation of the available information about the parameters. This row should be compared with the third row, which uses $T = T_{\text{true}} = 10$ (Eq. 15) in constructing the posteriors. The PPDs for the T obtained from the equipartition theorem and from T_{true} share the same general properties providing evidence that the equipartition theorem estimates capture the underlying nature of the PPDs.

An example of the PPDs where the estimate of T is too high is shown in the last row. For $T = 1000 >> T_{\text{true}}$, all of PPDs become uniformative. As T continues to increase, eventually the probability distributions all become uniform, which implies that the data contains no information about the parameters. This example highlights how the noise variance may be considered as a measure of the information content available to inform the inference of modeling parameters. Hence, properly matching the noise variance to the noise level of the data is important to obtain distributions which accurately represent the information content.

4. CONCLUSION

This work has shown that an analogy with the equipartition theorem in statistical mechanics can be used to obtain an estimate of the model temperature T for a simple data sample. With this estimate, the MaxEnt approach yields marginal PPDs that reflect the information content in the noisy data. Underestimation of T yields narrow, sharply peaked distributions appear with low variance. This implies greater information exists for the parameters than evidenced by the noisy data. Overestimation of T yields nearly flat PPDs for all parameters. The equipartition theorem estimate of T yields PPDs that have the same features as those obtained with the T value related to the standard deviation of the underlying Gaussian noise.

This method for estimating T from a single data sample is significant because prior work with the MaxEnt approach relied on estimates of T obtained from an ensemble over multiple data samples. The ensemble approach is limited to cases where multiple data samples are available, whereas, the equipartition theorem approach can be used *in situ* with only a single data sample.

This proof-of-concept example has used a Monte Carlo sampling approach. When sophisticated sampling methods can be used, the advantages of using the equipartition theorem estimate of T should still be evident.

ACKNOWLEDGMENTS

We acknowledge support from the Office of Naval Research Grant #N00014-22-12402 and thank the College of Physical and Mathematical Sciences at Brigham Young University for funding the first author's undergraduate research assistantship. We also express appreciation to the associate editor for their helpful review of the manuscript.

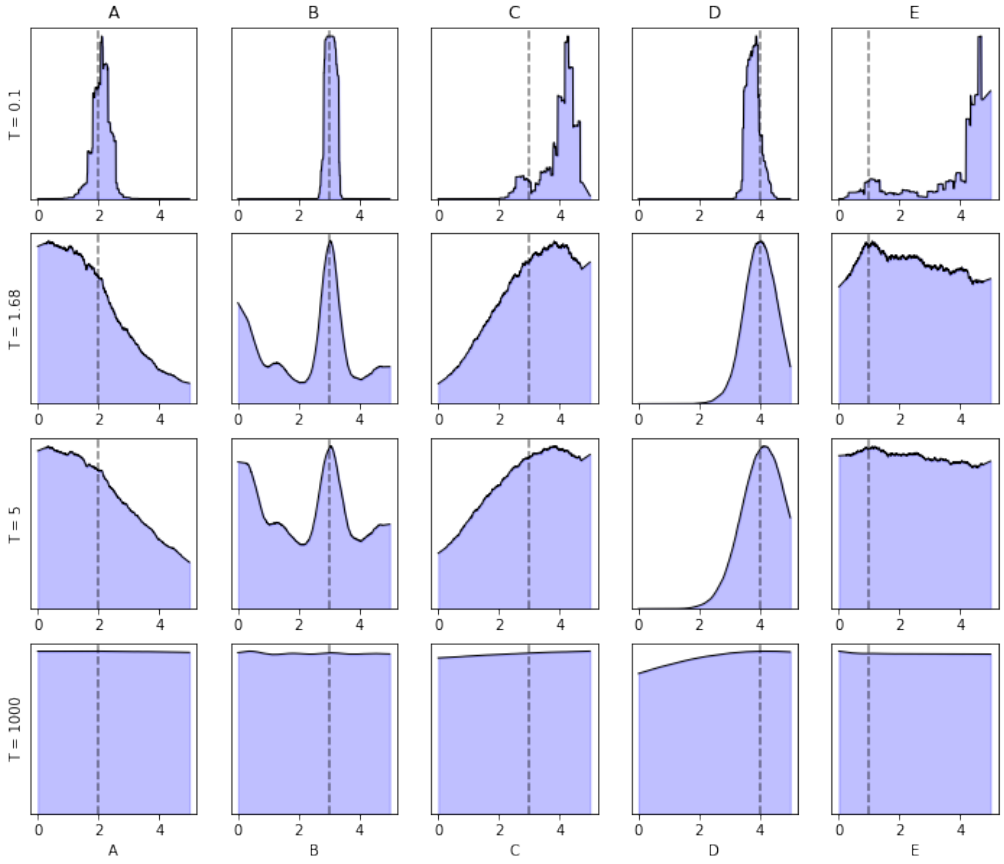

Figure 3: Examples of the impact of T on the PPDs obtained using the MaxEnt method on the toy model. True parameter values for each parameter are indicated by the dashed line. (Top row) Underestimation of the temperature of the model results in sharp, narrow distributions which imply more information than is available about the model from the data. (Bottom row) Overestimation of the temperature results in flatter, more uniform distributions which imply that there is less information than is available about the model from the data. (Second row) The temperature value $T = 1.68$ is obtained with Eq. 18. (Third row) The true temperature for this example is $T_{\mathrm{true}} = 2\sigma^2 = 10$ as in Eq. 15.

REFERENCES

[1] E. T. Jaynes, "Information theory and statistical mechanics," Phys. Rev. **106**(4), 620–630 (1957) 10.1103/PhysRev.106.620.

[2] E. T. Jaynes, "Information theory and statistical mechanics. part 2," Physical Review **108**(2), 171–190 (1957) 10.1103/PhysRev.108.171.

[3] E. T. Jaynes and G. L. Bretthorst, *Probability theory: the logic of science* (Cambridge University Press, Cambridge, UK; New York, NY, 2003).

[4] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal **27**(3), 379–423 (1948) 10.1002/j.1538-7305.1948.tb01338.x.

[5] E. T. Jaynes, "Prior probabilities," IEEE Transactions on Systems Science and Cybernetics **4**(3), 227–241 (1968) 10.1109/TSSC.1968.300117.

[6] E. T. Jaynes, *The Relation of Bayesian and Maximum Entropy Methods*, 25–29 (Springer Netherlands, Dordrecht), https://doi.org/10.1007/978-94-009-3049-0_2, 10.1007/978-94-009-3049-0_2.

[7] N. Xiang, "Model-based bayesian analysis in acoustics—a tutorial," The Journal of the Acoustical Society of America **148**(2), 1101–1120 (2020).

[8] J. van Campenhout and T. Cover, "Maximum entropy and conditional probability," IEEE Transactions on Information Theory **27**(4), 483–489 (1981) 10.1109/TIT.1981.1056374.

[9] P. M. Williams, "Bayesian conditionalisation and the principle of minimum information," The British Journal for the Philosophy of Science **31**(2), 131–144 (1980).

[10] E. T. Jaynes, *Probability theory: The logic of science* (Cambridge university press, 2003), pp. 355–357.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," The Annals of Mathematical Statistics **22**(1), 79–86 (1951).

[12] S. A. Stotts and R. A. Koch, "Application of maximum entropy to statistical inference for inversion of data from a single track segment," The Journal of the Acoustical Society of America **142**(2), 737–755 (2017) 10.1121/1.4996456.

[13] D. P. Knobles, J. D. Sagers, and R. A. Koch, "Maximum entropy approach to statistical inference for an ocean acoustic waveguide," The Journal of the Acoustical Society of America **131**(2), 1087–1101 (2012) 10.1121/1.3672709.

[14] D. P. Knobles, P. S. Wilson, J. A. Goff, L. Wan, M. J. Buckingham, J. D. Chaytor, and M. Badiey, "Maximum entropy derived statistics of sound-speed structure in a fine-grained sediment inferred from sparse broadband acoustic measurements on the new england continental shelf," IEEE Journal of Oceanic Engineering **45**(1), 161–173 (2020) 10.1109/JOE.2019.2922717.

[15] J. Gibbs, *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics*, Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics (C. Scribner's sons, 1902), https://books.google.com/books?id=2oc-AAAAIAAJ.

[16] A. Glazer and J. Wark, *Statistical Mechanics: A Survival Guide* (OUP Oxford, 2001), https://books.google.com/books?id=s_Ku5aUSq2IC.

[17] S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, "Bayesian ensemble approach to error estimation of interatomic potentials," Physical Review Letters **93**(16), 165501 (2004) 10.1103/PhysRevLett.93.165501.

[18] C. H. LaMont and P. A. Wiggins, "Correspondence between thermodynamics and inference," Physical Review E **99**(5), 052140 (2019) 10.1103/PhysRevE.99.052140.

[19] R. Christensen, T. Bligaard, and K. W. Jacobsen, *Bayesian error estimation in density functional theory*, 77–91 (Elsevier), https://linkinghub.elsevier.com/retrieve/pii/B9780081029411000031, 10.1016/B978-0-08-102941-1.00003-1.

[20] S. E. Dosso, "Quantifying uncertainty in geoacoustic inversion. i. a fast gibbs sampler approach," The Journal of the Acoustical Society of America **111**(1), 129–142 (2002).

[21] S. Dosso, P. Nielsen, and M. Wilmut, "Data error covariance in matched-field geoacoustic inversion," Journal of the Acoustical Society of America (2006).

[22] A. L. Bonomo and M. J. Isakson, "A comparison of three geoacoustic models using bayesian inversion and selection techniques applied to wave speed and attenuation measurements," The Journal of the Acoustical Society of America **143**(4), 2501–2513 (2018).

[23] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," Analytical Chemistry **36**(8), 1627–1639 (1964) 10.1021/ac60214a047.