M

Maximum Entropy Method

Dionissios T. Hristopulos[1] and Emmanouil A. Varouchakis[2]
[1]School of Electrical and Computer Engineering, Technical University of Crete, Chania, Greece
[2]School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece

Abbreviations

MEM	maximum entropy method
MaxEnt	maximum entropy
BME	Bayesian Maximum entropy

Definition

The principle of maximum entropy states that the most suitable probability model for a given system maximizes the Shannon entropy subject to the constraints imposed by the data and – if available – other prior knowledge of the system. The maximum entropy distribution is the most general probability distribution function conditionally on the constraints. In the geosciences, the principle of maximum entropy is mainly used in two ways: (1) in the maximum entropy method (MEM) for the parametric estimation of the power spectrum and (2) for constructing joint probability models suitable for spatial and spatiotemporal datasets.

Overview

The concept of *entropy* was introduced in thermodynamics by the German physicist Rudolf Clausius in the nineteenth century. Clausius used entropy to measure the thermal energy of a machine per unit temperature which cannot be used to generate useful work. The Austrian physicist Ludwig Boltzmann used entropy in statistical mechanics to quantify the *randomness (disorder)* of a system. The statistical mechanics definition of entropy reflects the number of microscopic configurations which are accessible by the system. In the twentieth century, the concept of entropy was used by the American mathematician Claude Shannon (1948) to measure the average information contained in signals. Shannon's influential paper founded the field of *information theory*. Consequently, the terms *Shannon entropy* and *information entropy* are used to distinguish between the entropy content of signals and the mechanistic notion of entropy used in thermodynamics and statistical mechanics.

The connection between information theory and statistical mechanics was investigated in two seminal papers by the American physicist Edwin T. Jaynes (1957a, b). He showed that the formulation of statistical mechanics can be derived from the principle of maximum entropy without the need for additional assumptions. The *principle of maximum entropy* is instrumental in establishing this connection; it dictates that given partial knowledge of the system, the least biased estimate for the probability distribution maximizes Shannon entropy under the specified constraints. According to Jaynes, "Entropy maximization is not an application of a law of physics, but merely a method of reasoning that ensures that no arbitrary assumptions are made." The work of Jaynes opened the door for the application of MEM to various fields of science and engineering that involved ill-posed problems characterized by incomplete information. Notable application areas include spectral analysis (Burg 1972), image restoration (Skilling and Bryan 1984), geostatistics (Christakos 1990), quantum mechanics, condensed matter physics, tomography, crystallography, chemical spectroscopy, and astronomy, among others (Skilling 2013).

B. S. Daya Sagar et al. (eds.), *Encyclopedia of Mathematical Geosciences*, Encyclopedia of Earth Sciences Series,
https://doi.org/10.1007/978-3-030-26050-7_196-1

Methodology

According to *Laplace's principle of indifference*, if prior knowledge regarding possible outcomes of an experiment is unavailable, the uniform probability distribution is the most impartial choice. MEM employs this principle by incorporating data-imposed constraints in the model inference process. The MEM probability model depends on the constraints used: The MEM model for a nonnegative random variable with known mean is the exponential distribution. If the constraints include the mean and the variance, the Gaussian (normal) distribution is obtained. In the case of multivariate and spatially distributed processes, the MEM model constrained on the mean and the covariance function is the joint Gaussian distribution.

Notation

In the following, it is assumed that $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ (T denotes the transpose) is an n-dimensional random vector defined in a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is the sample space, $\mathcal{F}$ is the sigma-algebra of events, and $\mathcal{P}$ is the probability measure. The realizations $\mathbf{x} \in \mathbb{R}^n$ of the random vector $\mathbf{X}$ can take either discrete or continuous values.

The expectation of the function $g(\mathbf{X})$ over the ensemble of states $\mathbf{x}$ is denoted by $\mathbb{E}[g(\mathbf{X})]$. The expectation involves the *joint probability mass function* (PMF) $p(\mathbf{x})$ if the random variables X_i take discrete values or the *joint probability density function* (PDF) $f(\mathbf{x})$ if the X_i are continuous. We use the *trace operator*, Tr, as a unifying symbol to denote summation (if the random variables X_i are discrete) or integration (if the X_i are continuous) over all probable states $\mathbf{x}$.

General Formulation

Assuming that the PMF $p(\mathbf{x})$ (in the discrete case) or the PDF $f(\mathbf{x})$ (in the continuous case) is known, Shannon's entropy can be expressed as

$$S = -\sum_{\mathbf{x}\in\Omega} p(\mathbf{x}) \ln p(\mathbf{x}), \quad \text{discrete}, \tag{1}$$

$$S = -\int_{\mathbb{R}} dx_1 \ldots \int_{\mathbb{R}} dx_n f(\mathbf{x}) \ln f(\mathbf{x}), \quad \text{continuous}. \tag{2}$$

The entropy for both the discrete and continuous cases can be expressed as $S = -\mathbb{E}[\ln f(\mathbf{x})]$, where $f(\cdot)$ here stands for the PMF in the discrete case and the PDF in the continuous case. For the sake of brevity, in the following, we use $f(\cdot)$ to denote the PDF or the PMF and refer to it as "probability distribution." We also use the term "summation" to denote either summation (for discrete variables) or integration (for continuous variables) over all possible states $\mathbf{x}$.

Let $\{g_m(\mathbf{x})\}_{m=1}^M$ represent a set of M sampling functions, the averages of which can be determined from the data (observations). We will denote sample averages by means of $\overline{g_m(\mathbf{x})}$. Such sampling averages can include the mean, variance, covariance, higher-order moments, or more complicated functions of the sample values. Then, according to the *principle of maximum entropy*, the probability distribution should respect the constraints (the symbol $\equiv$ denotes equivalence)

$$\mathbb{E}[g_m(\mathbf{x})] \equiv \mathrm{Tr}_{\mathbf{x}}[g_m(\mathbf{x}) f(\mathbf{x})] = \overline{g_m(\mathbf{x})}, \quad m = 1, \ldots, M, \tag{3}$$

where $\mathrm{Tr}_{\mathbf{x}}[\cdot]$ denotes the "summation" over all possible states $\mathbf{x}$. The above equations are supplemented by the *normalization constraint* $\mathrm{Tr}_{\mathbf{x}} f(\mathbf{x}) = 1$ which ensures the proper normalization of the probability distribution.

The maximum entropy (henceforward, *MaxEnt*) distribution maximizes the entropy under the above constraints. This implies a constrained optimization problem defined by means of the following Lagrange functional

$$\mathcal{L}[f] = -S + \sum_{m=1}^{M} \lambda_m \left[\overline{g_m(\mathbf{x})} - \mathrm{Tr}_{\mathbf{x}} g_m(\mathbf{x}) f(\mathbf{x})\right] + \lambda_0 [1 - \mathrm{Tr}_{\mathbf{x}} f(\mathbf{x})]. \tag{4}$$

The minimization of $\mathcal{L}[f]$ can be performed using the calculus of variations to find the stationary point of the Lagrange functional. At the stationary point, the functional derivative $\delta\mathcal{L}/\delta f$ vanishes, i.e.,

$$0 = \frac{\delta \mathcal{L}}{\delta f} = \ln f(\mathbf{x}) + 1 - \sum_{m=1}^{M} \lambda_m g_m(\mathbf{x}) - \lambda_0.$$

The above equation leads to the *MaxEnt probability distribution*

$$f(\mathbf{x}) = \frac{1}{Z} \exp\left[\sum_{m=1}^{M} \lambda_m g_m(\mathbf{x})\right], \quad \ln Z = 1 - \lambda_0, \tag{5}$$

where the constant Z is the so-called partition function which normalizes the MaxEnt distribution. Since $f(\mathbf{x})$ is normalized by construction, it follows from Eq. (5) that

$$Z = \mathrm{Tr}_{\mathbf{x}} \exp\left[\sum_{m=1}^{M} \lambda_m g_m(\mathbf{x})\right]. \tag{6}$$

The implication of Eqs. (5) and (6) is that λ_0 depends on the Lagrange multipliers $\{\lambda_m\}_{m=1}^M$. These need to be

determined by solving the following system of M nonlinear constraint equations

$$\overline{g_m(\mathbf{x})} = \mathrm{Tr}_{\mathbf{x}} g_m(\mathbf{x}) f(\mathbf{x}) = \frac{\partial Z}{\partial \lambda_m}, \quad m = 1, \ldots, M. \tag{7}$$

In spatial and spatiotemporal problems, the constraints involve joint moments of the field (Christakos 1990, 2000; Hristopulos 2020). The constraints are expressed in terms of real-space coordinates. In the case of time series analysis, it is customary to express the MEM solution in the spectral domain (see next section).

Spectral Analysis

MEM has found considerable success in geophysics as a method for estimating the power spectrum of stationary random processes (Burg 1972; Ulrych and Bishop 1975). In spectral analysis, MEM is also known as *all poles* and *autoregressive (AR) method*.

For a time series with a constant time step δt, the MEM power spectral density at frequency f is given by

$$P(f) = \frac{c_0}{\left|1 + \sum_{m=1}^{M} c_m \exp\left(2\pi i f m \delta t\right)\right|^2}, \quad -f_N \le f \le f_N, \tag{8}$$

where $f_N = 1/2\delta t$ is the Nyquist frequency (i.e., the maximum frequency which can be resolved with the time step δt) and $\{c_m\}_{m=0}^{M}$ are coefficients which need to be estimated from the data.

The term "all-poles" method becomes obvious based on Eq. (8), since $P(f)$ has poles in the complex plane. The poles coincide with the zeros in the denominator of the fraction that appears in the right-hand side of Eq. (8). The connection with AR time series models of order m lies in the fact that the latter share the spectral density given by Eq. (8). The order of the AR model which is constructed based on MaxEnt is equal to the maximum lag, $m\delta t$, for which the auto-covariance function can be reliably estimated based on the available data.

Applications

Maximum entropy has been extensively used in many disciplines of geoscience. Notable fields of application include geophysics, seismology and hydrology, estimation of rainfall variability and evapotranspiration, assessment of landslide susceptibility, classification of remote sensing imagery, prediction of categorical variables, investigations of mineral potential prospectivity, and applications in land use and climate models.

The principle of maximum entropy is used in Bayesian inference to obtain prior distributions (Skilling 2013). This connection has been exploited in the Bayesian MaxEnt (BME) framework for spatial and spatiotemporal model construction and estimation (Christakos 1990, 2000). BME allows incorporating prior physical knowledge of the spatial or spatiotemporal process in the probability model. In spatial estimation problems, the application of the method of maximum entropy assumes that the mean, covariance, and possibly higher-order moments provide the spatial constraints for random field models. A different approach proposes local constraints that involve geometric properties, such as the square of the gradient and the linearized curvature of the random field (Hristopulos 2020). This approach leads to spatial models with sparse precision matrix structure which is computationally beneficial for the estimation and prediction of large datasets. In recent years, various generalized entropy functions have been proposed (e.g., Ŕenyi, Tsallis, Kaniadakis entropies) for applications that involve long-memory, non-ergodic, and non-Gaussian processes. In principle, it is possible to build generalized MEM distributions using extended notions of entropy (Hristopulos 2020). The mathematical tractability and the application of such generalized maximum entropy principles in geoscience are open research topics.

Summary or Conclusions

The notion of entropy is central in statistical mechanics and information theory. Entropy provides a measure of the information incorporated in a specific signal or dataset. The principle of maximum entropy can be used to derive probability distributions which possess the largest possible uncertainty given the constraints imposed by the data. Equivalently, maximum entropy minimizes the amount of prior information which is integrated into the probability distribution. Thus, the maximum entropy probability distributions are unbiased with respect to what is not known. The most prominent applications of maximum entropy in geoscience include the estimation of spectral density for stationary random processes and the construction of joint probability models for spatial and spatiotemporal datasets.

Cross-References

- Discrete Fourier Transform
- Higher-Order Spatial Stochastic Models
- Power Spectral Density
- Spatial Analysis

- Spatial Statistics
- Spectral Analysis
- Stochastic Process
- Time Series Analysis

Bibliography

Burg JP (1972) The relationship between maximum entropy spectra and maximum likelihood spectra. Geophysics 37(2):375–376

Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. Math Geol 22(7):763–777

Christakos G (2000) Modern spatiotemporal geostatistics, International Association for Mathematical Geology Studies in mathematical geology, vol 6. Oxford University Press, Oxford

Hristopulos DT (2020) Random fields for spatial data modeling: a primer for scientists and engineers. Springer Netherlands, Dordrecht

Jaynes ET (1957a) Information theory and statistical mechanics. I. Phys Rev 106(4):620–630

Jaynes ET (1957b) Information theory and statistical mechanics. II. Phys Rev 108(2):171–190

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Skilling J (2013) Maximum entropy and Bayesian methods: Cambridge, England, 1988, vol 36. Springer Science & Business Media, Cham

Skilling J, Bryan R (1984) Maximum entropy image reconstruction-general algorithm. Mon Not R Astron Soc 211:111–124

Ulrych TJ, Bishop TN (1975) Maximum entropy spectral analysis and autoregressive decomposition. Rev Geophys 13(1):183–200