J. Geophys. 42, 489–506, 1977

Journal of Geophysics

The Maximum Entropy Approach to Inverse Problems

Spectral Analysis of Short Data Records and Density Structure of the Earth *

E. Rietsch

Deutsche Texaco Aktiengesellschaft, Hauptlaboratorium für Erdölgewinnung,
D-3101 Wietze/Celle, Federal Republic of Germany

Abstract. The maximum entropy principle as described in the first, introductory part of the paper is applied to 2 problems: the estimation of the power spectrum from a finite number of values of the autocovariance function, and the determination of the density within the Earth from its mass, radius, and moment of inertia. In both cases the available information is given in terms of known values of linear functionals and the maximum entropy principle is used to derive a probability distribution for the values of the unknown function. The expectation value of the probability distribution for the spectral power is shown to be equal to the well-known maximum entropy power spectrum. The expectation value for the density within the Earth is in – with respect to the few data used – good agreement with that of accepted Earth models.

Key words: Maximum entropy – Probability distribution – Inverse problem – Power spectrum – Density – Earth.

Introduction

The problem of estimation of a large number of unknowns or an unknown function from only few measured values of functionals of these unknowns constitutes a typical task in geophysics. A deterministic approach to its solution is, for example, achieved by the well-known Backus-Gilbert inversion technique (Backus and Gilbert, 1967, 1968, 1970).

In this paper it is proposed to handle such problems by means of probabilistic methods based on the maximum entropy principle as put forward by Jaynes in 1968.

The paper starts with a review of the ideas which lead to the formulation of this principle. This introductory part is of a more tutorial character and intends

[*] Edited version of a paper presented at the Institute for Geophysics, Technical University Clausthal, on June 10, 1976

to explain the reasons for maximizing the entropy of a probability distribution. The equations established in this part are then applied to two examples. In the first a probability distribution for the spectral power of a band-limited time series is derived from the first few values of its autocovariance function. The expectation value of this probability distribution is shown to be the well known maximum entropy spectrum. In the second example a probability distribution is derived for the density as a function of depth of a spherically symmetric Earth assuming radius, mass, and moment of inertia to be known. The expectation value of the density obtained from of this probability distribution agrees amazingly well with the, according to Bullen (1975), most likely density distribution.

The Concept of Entropy

The Entropy of Discrete Probability Distributions

In probability theory a finite, complete system of events is understood to mean a set of events

$$A_1, A_2, \ldots, A_n$$

such that, as a result of a certain experiment, one and only one of these events must occur. To each of these events A_k there is associated a probability of occurrence $p_k \geqq 0$. The system of events and the associated probabilities may be arranged in the scheme

$$A = \begin{pmatrix} A_1 & A_2 & \ldots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix},$$

and the completeness of the scheme A is expressed by

$$\sum_{j=1}^{n} p_j = 1. \tag{1}$$

The tossing of a coin, for example, may be described by the scheme

$$A = \begin{pmatrix} A_1 & A_2 \\ 1/2 & 1/2 \end{pmatrix},$$

which consists of the two events A_1 and A_2 representing the two possible outcomes heads and tails.

Any such scheme describes a situation of uncertainty. One knows that an experiment will lead to one of n possible events but is unable to predict with certainty which of these events will eventually occur.

Obviously, the amount of uncertainty concerning the outcome of an experiment is different in different schemes. Consider the scheme

$$B = \begin{pmatrix} B_1 & B_2 \\ \dfrac{1}{1024} & \dfrac{1023}{1024} \end{pmatrix}$$

which describes the simultaneous tossing of ten coins. The event B_1 occurs when all ten coins show heads, and the event B_2 comprises all other possible combinations of heads and tails for these ten coins. Clearly, in this scheme there is much less uncertainty than in scheme A. An experimenter will be almost sure to have the event B_2 as outcome of his experiment whereas he would refrain from any prediction in the situation described by scheme A.

Realization of a given scheme, i.e. performing of the experiment the possible outcomes (events) of which are described by this scheme, completely removes the uncertainty. Hence the average information obtained by carrying out the experiment (namely the information which of the possible events actually occurred) may be regarded as proportional to the uncertainty that existed before the experiment. Sometimes the notion "average information" is used synonymously with "uncertainty".

In this situation it seems desirable to have a measure for the amount of uncertainty inherent in a particular scheme or, equivalently, a measure of the average information obtained from a realization of this scheme.

Such a measure – which will be a function of the probabilities of the different events – must satisfy a number of reasonable consistency requirements (Aczél and Daróczy, 1975). From among these conditions the following three are sufficient to define a function $H(\mathbf{p})=H(p_1,\ldots,p_n)$ which serves this purpose and is unique – apart from a positive constant factor (Khinchin, 1957).

a) *The uncertainty asslciated with a finite complete scheme A takes its largest value if all events are equally likely.*

Because of Equation (1) this means

$$H(p_1,\ldots,p_n)\leqq H(1/n,\ldots,1/n). \qquad (2)$$

b) *Addition of an impossible event to a scheme does not change the amount of uncertainty.*

The amount of uncertainty is therefore equal in the two schemes

$$\begin{pmatrix} A_1 & A_2 & \ldots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} A_1 & A_2 & \ldots & A_n & A_{n+1} \\ p_1 & p_2 & \cdots & p_n & 0 \end{pmatrix}.$$

In terms of the function $H(\mathbf{p})$ this condition reads

$$H(p_1,\ldots,p_n)=H(p_1,\ldots,p_n,0). \qquad (3)$$

c) *The uncertainty in the product AB of the two schemes A and B is equal to the uncertainty in scheme A increased by the uncertainty remaining in scheme B after a realization of scheme A.*

Alternatively, be means of the previously introduced notion of the "average information", this third condition may be expressed as follows: The average information obtained from a realization of scheme AB is equal to the average information obtained from a realization of scheme A increased by the additional average information expected from a realization of scheme B after realization of scheme A.

The meaning of this condition is explained in the following. Let A and B denote 2 finite schemes with n and m possible events, respectively. The product scheme AB consists of the $n\,m$ combinations $A_j B_k$ of events. Let p_j denote the probability of event A_j in scheme A and let q_{jk} stand for the (conditional) probability that the event B_k of scheme B occurs provided that the event A_j in scheme A occurred. Then the product scheme AB has the form

$$AB=\begin{pmatrix} A_1B_1 & A_1B_2 & \dots & A_1B_m & A_2B_1 & \dots & A_nB_m \\ p_1q_{11} & p_1q_{12} & \cdots & p_1q_{1m} & p_2q_{21} & \cdots & p_nq_{nm} \end{pmatrix}$$

and condition c) demands that

$$H(AB)=H(A)+H(B|A). \tag{4}$$

The abbreviations

$$H(AB)=H(p_1q_{11},\dots,p_nq_{nm}), \qquad H(A)=H(p_1,\dots,p_n)$$

denote the uncertainty in schemes AB and A, respectively, and

$$H(B|A)=\sum_j p_j H(q_{j1},\dots,q_{jm})$$

describes the uncertainty in scheme B after a realization of scheme A. $H(q_{j1},\dots,q_{jm})$ denotes the uncertainty in scheme B after occurrence of event A_j in scheme A. This function is multiplied with the probability p_j of event A_j and summed over all possible events of scheme A to obtain $H(B|A)$.

If the two schemes A and B are independent a realization of scheme A does not convey any information concerning the outcome of a realization of scheme B: the q_{jk} do not depend on j. In this case $H(B|A)=H(B)$.

On the other hand, $H(B|A)=0$ if the outcome of scheme A completely determines the outcome of scheme B.

The above three conditions lead to a function (Khinchin, 1957)

$$H(p_1,\dots,p_n)=-\lambda\sum_{j=1}^{n} p_j\log p_j \tag{5}$$

which is called "entropy" in view of an analogy with the entropy in thermodynamics. It is unique apart from the positive factor λ which frequently is set to 1/log 2. Hence it can be omitted if the logarithm is to the base 2 as done in the following.

It is easy to show that the function H defined in (5) satisfies the above stated three basic conditions.

Another obviously necessary property of the uncertainty in a finite scheme can also be easily established from (5): the entropy is zero if and only if one of the numbers p_j is unity and all others are zero. This is just the case when the outcome of an experiment can be predicted with certainty.

Incorporation of Information

The first of the three conditions which lead to the mathematical expression for the entropy specified that the entropy must have a maximum when all possible outcomes of an experiment are equally likely. Such a situation exists if there is no reason to consider one possible outcome of an experiment to be more likely than any other. This condition is related to Bernoulli's principle of insufficient reason or Keynes' principle of indifference (Rowlinson, 1970). It is a subjective principle and does not necessarily mean that all events are really equally probable but only that one's state of knowledge is not sufficient to assign to some of the outcomes a higher probability than to others. However, if one is given information concerning the outcome of the experiment it should be exploited to obtain better estimates of the probabilities. The foregoing analysis gives a clue how this can be accomplished. The distribution of probabilities among the possible events should maximize the uncertainty of the scheme without contradicting the given information. Any other choice of probabilities would either lead to a probability distribution with lower entropy – thus implying that further information has been assumed – or contradict the available information or both.

Let this information be given in the form of mean values $\bar{f}_1, \ldots, \bar{f}_m$ of m functions $f_1(A_j), \ldots, f_m(A_j)$ of the n possible outcomes A_j, where $m<n$. The distribution of probabilities complying with this information and free from all other assumptions is the one which maximizes the entropy

$$H = -\sum_j p_j \log p_j \tag{6}$$

subject to the constraints

$$\sum_j p_j = 1 \tag{7}$$

$$\sum_j p_j f_k(A_j) = \bar{f}_k \qquad k=1,2,\ldots,m. \tag{8}$$

Here and in the following the summations over j and k are understood to extend from 1 to n and 1 to m, respectively. Condition (7) means that the system of events is complete while the m equations (8) specify the known mean values $\bar{f}_k$ of the functions $f_k(A_j)$. The solution to this problem by means of Lagrange multipliers is straightforward and leads to

$$p_j = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\sum_k \lambda_k f_k(A_j)\right]. \tag{9}$$

The form of the partition function

$$Z(\boldsymbol{\lambda}) = Z(\lambda_1, \ldots, \lambda_m) = \sum_j \exp\left[\sum_k \lambda_k f_k(A_j)\right] \tag{10}$$

makes immediately clear that the normalization condition (7) is satisfied. The m Lagrange multipliers λ_k are to be determined from the m conditions (8) which can be brought into the form

$$\frac{\partial}{\partial \lambda_k} \ln Z(\boldsymbol{\lambda}) = \bar{f}_k. \tag{11}$$

The entropy reads

$$H = \left(\ln Z(\boldsymbol{\lambda}) - \sum_k \lambda_k \bar{f}_k\right)/\ln 2. \tag{12}$$

Entropy of Continuous Probability Distributions

The derivation of the entropy (5) as a measure of the uncertainty is only valid for discrete probability distributions. This represents a limitation and, in view of the success of the maximum entropy principle in discrete cases, makes it desirable to extend the concept of entropy to random variables for which a continuum of values is permitted.

Let X denote a random variable which may have values x in some interval $[a, b]$ and let $p(x)\,dx$ denote the probability that a value of X be in the interval[1] $[x, x+dx]$. A selfsuggesting approach to the definition of the entropy of this continuous probability distribution is to subdivide the interval $[a, b]$ into subintervals $[x_{j-1}, x_j]$ where

$$a = x_0 < x_1 < \cdots < x_n = b$$

and to denote by $p_j \Delta x_j$ the probability that x is in the interval $[x_{j-1}, x_j]$ of length $\Delta x_j = x_j - x_{j-1}$. The entropy of this discretized probability distribution reads

$$\begin{aligned} H &= -\sum_j p_j \Delta x_j \log(p_j \Delta x_j) \\ &= -\sum_j p_j \Delta x_j \log p_j - \sum_j p_j \Delta x_j \log \Delta x_j . \end{aligned} \tag{13}$$

For $n \to \infty$ and $\max(\Delta x_j) \to 0$ the first term to the right of the last equality sign in (13) passes to

$$\int p(x) \log p(x)\, dx$$

where the integral is understood to extend over the interval $[a, b]$. The second term requires special analysis. By means of weights w_j defined as

$$\Delta x_j = \delta / w_j \qquad \sum_j w_j \Delta x_j = n\,\delta = 1 \tag{14}$$

and the discretized normalization condition

$$\sum_j p_j \Delta x_j = 1$$

it can be rewritten as

$$-\sum_j p_j \Delta x_j \log \Delta x_j = \sum_j p_j \Delta x_j \log w_j + \log n .$$

By means of this procedure the last term in (13) has been separated into a finite term which approaches the Riemann integral

$$\int p(x) \log w(x)\, dx$$

for $n \to \infty$ and $\max(\Delta x_j) \to 0$ and a divergent term which, however, does not depend on the particular subdivision of the interval $[a, b]$. Neglecting of this divergent term leads to the following equation for the entropy of a continuous probability distribution

$$H = -\int p(x) \log[p(x)/w(x)]\, dx \tag{15}$$

[1] In order to avoid unnecessary complications open and closed ends of intervals are not distinguished

which depends not only on $p(x)$ but also on $w(x)$, an "invariant measure" function (Jaynes, 1968) proportional to the varying density of the x_j in the limiting case.

Incorporation of Information

Let the information about the probability distribution $p(x)$ again be given in form of known mean values $\bar{f}_k$ of m different functions $f_k(x)$.

$$\int p(x) f_k(x)\,dx = \bar{f}_k \qquad k = 1, \ldots, m. \tag{16}$$

Maximization of H subject to these conditions leads to

$$\begin{aligned} p(x) &= [w(x)/Z(\lambda)] \exp\left[\sum_k \lambda_k f_k(x)\right] \\ Z(\lambda) &= \int w(x) \exp\left[\sum_k \lambda_k f_k(x)\right] dx \end{aligned} \tag{17}$$

and the Lagrange multipliers λ_k are determined by means of Equation (11).

At this stage there appears a practical difficulty not present in the discrete case discussed before. The measure $w(x)$ defining the distribution of the x_j in the limiting case is as yet undetermined unless there exists an obvious limiting process. The meaning of $w(x)$ becomes clear if a situation is considered in which no prior information is available. In this case there are no Lagrange multipliers and from Equations (17) follows

$$p(x) = w(x) \tag{18}$$

by virtue of

$$\int w(x)\,dx = 1 \tag{19}$$

obtained from (14) by passing to the limit $n \to \infty$, $\max(\Delta x_j) \to 0$. Hence $w(x)$ represents the prior probability distribution existing in the case of complete ignorance. Since the Lagrange multipliers may only be determined if $w(x)$ is known the question arises how to find this prior probability distribution.

A uniform prior probability distribution can not generally be appropriate since it is not invariant under coordinate transformations *and* since there exists apparently no general criterion for finding a "distinguished" coordinate system.

Jaynes (1968) proposed to use group theoretical reasoning, in particular to search for transformations under which there is no change in the level of ignorance and to require that $w(x)$ be invariant with respect to these transformations. This approach has been used to determine the prior probability distribution in the first of the following 2 examples.

Another possibility has already been mentioned. As often is the case with physical quantities, the variable X may be continuous as a result of an abstraction process; i.e., in principle, only a discrete set of values x_j is permissible for X with the difference $\Delta x_j = x_j - x_{j-1}$, however, so small that, for practical applications, X may be assumed to be continuous. Then there exists an obvious discretization and $w(x)$ can be determined. This possibility is employed in the second example.

Multivariate Probability Distributions

Generalization to multivariate probability distributions of Equations (15) to (17) is straightforward. Let $\mathbf{X}$ denote a vector of random variables $(X_1, \ldots, X_n)$ and let $p(\mathbf{x})$ denote the multivariate (or joint) probability distribution $p(x_1, \ldots, x_n)$. With $w(\mathbf{x})$ representing the multivariate prior probability distribution the entropy reads

$$H = -\int p(\mathbf{x}) \log[p(\mathbf{x})/w(\mathbf{x})]\, dv. \tag{20}$$

Here and in the following dv denotes the n-dimensional volume element, and the integration is extended over all possible values of the random variables X_j.

If the information about $\mathbf{X}$ is given in the form

$$\int p(\mathbf{x}) f_k(\mathbf{x})\, dv = \bar{f}_k \tag{21}$$

multivariate probability distribution and partition function are given by

$$\begin{aligned} p(\mathbf{x}) &= [w(\mathbf{x})/Z(\boldsymbol{\lambda})] \exp\Big[\sum_k \lambda_k f_k(\mathbf{x})\Big] \\ Z(\boldsymbol{\lambda}) &= \int w(\mathbf{x}) \exp\Big[\sum_k \lambda_k f_k(\mathbf{x})\Big]\, dv. \end{aligned} \tag{22}$$

Spectral Analysis of Time Series by Means of the Autocovariance Function

Statement of the Problem

Let

$$A_n = \int P(f) \exp(2\pi i f n \Delta t)\, df \tag{23}$$

denote the autocovariance function of a time series sampled at equidistant time values $n\Delta t$. The upper frequency limit be equal to the Nyquist frequency $f_\nu = 1/(2\Delta t)$ and the integral in (23) is understood to extend from $-f_\nu$ to f_ν. For a real-valued time series the A_n are real and symmetric and $P(f)$ is real, symmetric, and nonnegative.

It is required to estimate $P(f)$ from a limited number of values of the autocovariance function.

Usually this is accomplished by invoking the Fourier inversion theorem. Owing to the finite number of available values of the autocovariance function only a smoothed version

$$\tilde{P}(f) = \Delta t \sum_n A_n W_n \exp(-2\pi i f n \Delta t)$$

of the true power spectrum $P(f)$ may be obtained (Kanasewich, 1975)

$$\tilde{P}(f) = \int K(f - f') P(f')\, df'.$$

The amount of smoothing can be influenced by proper selection of the coefficients W_n of the Fourier kernel

$$K(f) = \Delta t \sum_n W_n \exp(-2\pi i f n \Delta t).$$

A formally identical result is obtained if the problem is attacked by the Backus-Gilbert inversion technique which leads to a particular kernel (Backus and Gilbert, 1968, appendix B). Kernels of this type have been investigated by Papoulis (1973).

It is the objective of this example to show how this problem can be handled by means of the above derived maximum entropy equations. For this purpose Equation (23) is considered as the limiting case of the periodic autocovariance function A_n^T with period T.

$$A_n = \lim_{T\to\infty} A_n^T = \lim_{M\to\infty} \left[\Delta f \sum_m P_m \exp(2\pi\, i\, n\, m/M)\right]. \tag{24}$$

Here

$$\Delta f = 1/T, \qquad P_m = P(m/T), \qquad M = 2f_v T \gg 2N.$$

The summation over m in Equation (24) extends from $[-(M-1)/2]$ to $[(M-1)/2]$ or from 0 to $M-1$ (here brackets denote the greatest integer less or equal to the enclosed expression). This reduces the problem to the determination of the spectral power at discrete frequencies $f_m = m/T$.

Since the number of unknowns grossly exceeds the number of known values of the autocovariance function no attempt is made to determine a fixed value of the spectral power at each of the discrete frequencies. Instead, a probability distribution of the spectral power is derived. In the spirit of the foregoing analysis it is requested that this probability distribution maximizes the entropy while accounting for the information available in form of the known values of the autocovariance function. Hence the maximum entropy principle may be formulated as follows.

Maximize the entropy

$$H = -\int p(\mathbf{P}) \log[p(\mathbf{P})/w(\mathbf{P})]\, dv \tag{25}$$

of the multivariate probability distribution $p(\mathbf{P}) = p(P_{[-(M-1)/2]}, \ldots, P_{[(M-1)/2]})$ subject to the normalization condition

$$\int p(\mathbf{P})\, dv = 1 \tag{26}$$

and

$$\Delta f \int p(\mathbf{P}) \left[\sum_m P_m \exp(2\pi\, i\, n\, m/M)\right] dv = \bar{A}_n \qquad n = -N, \ldots, N. \tag{27}$$

The last condition implies that the expectation value of the Fourier transform of the power spectrum should agree with the known values $\bar{A}_n$ of the autocovariance function. The integrations in Equations (25), (26) and (27) are performed over all positive values of $P_{[-(M-1)/2]}, \ldots, P_{[(M-1)/2]}$, and dv denotes the M-dimensional volume element.

The resulting multivariate probability distribution has the form (Eq. (22))

$$p(\mathbf{P}) = [w(\mathbf{P})/Z(\lambda)] \exp\left(-\sum_m L_m P_m\right) \tag{28}$$

$$Z(\lambda) = \int w(\mathbf{P}) \exp\left(-\sum_m L_m P_m\right) dv \tag{29}$$

where

$$L_m = \sum_n \lambda_n \exp(2\pi\, i\, n\, m/M) \tag{30}$$

and λ_n denote the $2N+1$ Lagrange multipliers.

Here and in the following summation over n is understood to extend from $-N$ to N while summation over m goes from $[-(M-1)/2]$ to $[(M-1)/2]$.

Prior Probability Distribution of the Power Spectrum

Determination of the Lagrange multipliers requires knowledge of $w(\mathbf{P})$. This prior probability distribution of the spectral power can be determined on the basis of the following 3 conditions.

1. *The spectral power $P(f)$ can be represented as the sum of the squares of the Fourier cosine component $C(f)$ and the Fourier sine component $S(f)$.*

$$P(f) = C^2(f) + S^2(f).$$

2. *$C(f)$ and $S(f)$ are independent from each other, i.e. knowledge of one of the Fourier components conveys no information concerning the value of the other Fourier component.*

3. *The probability distribution of $C(f)$ and $S(f)$ does not depend on the origin of the time axis.*

This last condition means that the new Fourier coefficients

$$C'(f) = \cos\varphi\; C(f) - \sin\varphi\; S(f)$$
$$S'(f) = \sin\varphi\; C(f) + \cos\varphi\; S(f)$$

which are obtained in place of $C(f)$ and $S(f)$ if the origin of the time axis is shifted by τ should have the same probability distribution—for any $\varphi = 2\pi f\tau$.

The joint probability distribution for C and S (for the sake of simplicity the argument f is omitted) is, because of condition 2, equal to the product $p_c(C)\, p_s(S)$ of the probability distribution of the individual Fourier components. Condition 3 requires that

$$p_c(C)\, p_s(S) = p_c(\cos\varphi\; C - \sin\varphi\; S)\, p_s(\sin\varphi\; C + \cos\varphi\; S). \tag{31}$$

Putting $\varphi = \pi/2$ one finds

$$p_c(C)/p_s(C) = p_c(-S)/p_s(S)$$

for any C and S. Hence the two Fourier components must have the same probability distribution.

Let $y = \cos\varphi\; C$, $z = \sin\varphi\; C$, $S = 0$. Then Equation (31) reduces to the well known functional equation

$$p_c(\sqrt{y^2 + z^2})\, p_c(0) = p_c(y)\, p_c(z)$$

which has the normalized solution (Rao, 1965)

$$p_c(C) = \sqrt{\lambda/\pi}\exp(-\lambda C^2) \tag{32}$$

with the arbitrary positive constant λ. Because of condition 1 the probability distribution for the spectral power derived from Equation (32) reads

$$w(P) = \lambda\exp(-\lambda P). \tag{33}$$

If knowledge of the power at one frequency does not convey information concerning the power at any other frequency the multivariate prior probability distribution is the product of the prior probability distributions of the spectral power at the frequencies f_m. Furthermore the parameter λ should be equal for all frequencies.

Hence

$$w(\mathbf{P}) = \lambda^M \exp(-\lambda \sum_m P_m). \tag{34}$$

Probability Distribution and Expectation Value of the Spectral Power

The parameter λ in Equation (34) plays the role of a scaling factor and may be included into the Lagrange multiplier λ_0 in L_m (Eqs. (28) and (30)). The multivariate probability distribution $p(\mathbf{P})$ factors into the product of the probability distributions for the individual frequencies

$$p(\mathbf{P}) = \prod_m p_m(P_m)$$

with

$$p_m(P_m) = L_m \exp(-L_m P_m). \tag{35}$$

The expectation value of the spectral power at frequency f_m reads

$$\bar{P}_m = \int P_m\, p_m(P_m)\, dP_m = 1/L_m. \tag{36}$$

At this stage it is convenient to pass to the limit $T \to \infty$. Then

$$L_m \to L(f) = \sum_n \lambda_n \exp(2\pi i f n \Delta t) \tag{37}$$

$$\bar{P}_m \to \bar{P}(f) = 1/L(f). \tag{38}$$

In order to determine the Lagrange multipliers λ_n we note that the power spectrum must be positive and integrable. Hence $L(f)$ must be nonnegative and, by the Fejér-Riesz Theorem (Akhiezer, 1956), allows factoring

$$L(f) = \lambda G(f)\, G^*(f) = \lambda |G(f)|^2$$

where

$$G(f) = \sum_{\nu=0}^{N} g_\nu \exp(2\pi i f \nu \Delta t),$$

and λ denotes a factor chosen to make $g_0 = 1$.

Multiplication of both sides of (38) by $G^*(f)\exp(2\pi i f n \Delta t)$ and integration from $-f_v$ to f_v leads to (Edward and Fitelson, 1973)

$$\sum_{\nu=0}^{N} \bar{A}_{n-\nu}\, g_\nu = \frac{1}{\lambda \Delta t}\, \delta_{n0} \tag{39}$$

where $\bar{A}_n$ denote the given values of the autocovariance function and δ_{n0} the Kronecker symbol. From this linear system of equations the g_ν can be identified as the coefficients of the $(N+1)$-length prediction error filter with $P_{N+1} = 1/(\lambda\, \Delta \mathrm{t})$ being the power of the unpredictable noise. Thus the expectation value $\bar{P}(f)$ of the probability distribution for the spectral power is equal to the well known maximum entropy power spectrum

$$\bar{P}(f) = 1/[\lambda |G(f)|^2] = P_{N+1}\, \Delta t / |\sum_{\nu} g_\nu \exp(2\pi i f \nu \Delta t)|^2 \tag{40}$$

which has found application in many different branches of geophysics (e.g. Ulrych, 1972; Smylie, Clarke and Ulrych, 1973; Jensen and Ulrych, 1973; Phillips and Cox, 1976).

Density Distribution within a Spherically Symmetric Earth

Statement of the Problem

This second example to be discussed refers to the following question: What can be said about the density distribution $\rho(r)$ within a (spherically symmetric) planet if its mass M, radius R, and moment of inertia J are known (in order to have specific data this planet is assumed to be the Earth).

This problem which constitutes part of an investigation by Cook (1971) could, in principle, be attacked by the Backus-Gilbert inversion method which would result in a smoothed version of the true density distribution. Here, however, the data are so inadequate with respect to the desired solution that the "resolution length" which characterizes the amount of smoothing is of the order of R and thus renders the results rather useless.

Parker (1972) has adopted another approach to this problem. He proposed to look for inequalities resulting from these data such that all Earth models complying with the given data satisfy the inequalities. In this way he concluded, for example, that the maximum density in the Earth must not be less than ρ_0 where

$$\rho_0 = \bar{\rho}/y^{3/2} = 1.299\, \bar{\rho} = 7.166\ \mathrm{g/cm^3} \tag{41}$$

and

$$\bar{\rho} = 5.517\ \mathrm{g/cm^3}, \qquad y = 5J/(2MR^2) = 0.84 \tag{42}$$

denote mean density (Bullen, 1975) and a dimension less factor proportional to the ratio of the actual moment of inertia of the Earth and the moment of inertia of a homogeneous sphere with the same mass and radius, respectively.

This bound has much in common with error bounds e.g. in numerical integration, summation, or inversion of matrices. These bounds are safe but usually overly pessimistic; in general, the error is much lower than indicated by the bounds.

This example is to demonstrate the maximum entropy approach to this problem. To this aim the Earth is subdivided into N equivoluminous shells and to each shell there is assigned a constant density ρ_n. By means of this discretization

$$M=4\pi\int_0^R \rho(r)\, r^2\, dr \to \frac{4\pi}{3}\sum_n \rho_n(r_n^3-r_{n-1}^3) \tag{43}$$

$$J=\frac{8\pi}{3}\int_0^R \rho(r)\, r^4\, dr \to \frac{8\pi}{15}\sum_n \rho_n(r_n^5-r_{n-1}^5) \tag{44}$$

with

$$r_n=R(n/N)^{1/3} \tag{45}$$

denoting the outer radius of the n-th shell. Here and in the following the summation over n is understood to extend from 1 to N.

For $N>2$ it is impossible to uniquely determine the ρ_n and hence a probability distribution $p(\boldsymbol{\rho})$ is established for the densities of the N shells. This multivariate probability distribution should maximize the entropy

$$H=-\int p(\boldsymbol{\rho}) \log[p(\boldsymbol{\rho})/w(\boldsymbol{\rho})]\, dv \tag{46}$$

subject to condition

$$\int p(\boldsymbol{\rho})\, dv=1. \tag{47}$$

Furthermore, the expectation values

$$\bar{\rho}_n=\int \rho_n\, p(\boldsymbol{\rho})\, dv \tag{48}$$

should satisfy

$$\sum_n \bar{\rho}_n=\bar{\rho}\,\mathrm{N} \tag{49}$$

$$\sum_n \bar{\rho}_n[n^{5/3}-(n-1)^{5/3}]=y\,\bar{\rho}\,N^{5/3}. \tag{50}$$

Prior Probability Distribution

Before proceeding further we shall determine the appropriate prior probability distribution for the densities ρ_n. In this case it is easiest to look for a physically reasonable discretization.

In view of the atomistic nature of matter the density of a pure material is proportional to the number of molecules per volume element. Therefore the density can have values from a discrete set only. Adding or removing one molecule from the volume element changes the density by a fixed constant amount. Thus the range of possible values for the density is to be subdivided into intervals of equal width and the concentration of subdivision points is a constant. Hence a constant prior probability distribution is appropriate.

This line of argument is not invalidated if different types of molecules are permitted. In this case the density can change in different steps, and as long as the change in density is independent from the number of molecules already present the limiting distribution of subdivision points is still a constant. Since the Earth has been subdivided into shells of equal volume the prior probability distribution is the same for all shells.

Probability Distribution and Expectation Value for the Density

Maximization of the entropy (46) subject to Equations (47), (49) and (50) leads to

$$p(\boldsymbol{\rho})=\exp\left[-\sum_n \rho_n h_n(\boldsymbol{\lambda})\right]/Z(\boldsymbol{\lambda}) \tag{51}$$

where

$$h_n(\boldsymbol{\lambda})=\lambda_1+\lambda_2\left[n^{5/3}-(n-1)^{5/3}\right]/N^{2/3} \tag{52}$$

$$Z(\boldsymbol{\lambda})=\int \exp\left[-\sum_n \rho_n h_n(\boldsymbol{\lambda})\right] dv\,. \tag{53}$$

From (51) follows that the multivariate probability distribution $p(\boldsymbol{\rho})$ is the product of the probability distributions

$$p_n(\rho_n)=\exp\left[-\rho_n h_n(\boldsymbol{\lambda})\right]/Z_n(\boldsymbol{\lambda}) \tag{54}$$

of the densities of the N shells.

Let $\rho_l \geqq 0$ and $\rho_u \leqq N\bar{\rho}$ denote the lowest and highest possible density, respectively. Then by (53) and (54)

$$Z_n(\boldsymbol{\lambda})=\exp\left[-\rho_l h_n(\boldsymbol{\lambda})\right]\left[1-1/E_n(\boldsymbol{\lambda})\right]/h_n(\boldsymbol{\lambda}) \tag{55}$$

with

$$E_n(\boldsymbol{\lambda})=\exp\left[(\rho_u-\rho_l)\, h_n(\boldsymbol{\lambda})\right], \tag{56}$$

and the expectation value $\bar{\rho}_n$ has the form

$$\bar{\rho}_n=\int \rho_n p_n(\rho_n)\, d\rho_n=\rho_l+1/h_n(\boldsymbol{\lambda})-(\rho_u-\rho_l)/\left[E_n(\boldsymbol{\lambda})-1\right]. \tag{57}$$

The unknown Lagrange multipliers λ_1, λ_2 can be obtained from Equations (49) and (50) with $\bar{\rho}_n$ substituted from Equation (57).

For the upper density limit ρ_u sufficiently large (say $\rho_u \geqq 300$ g/cm^3) the third term in Equation (57) may be neglected. In this case Equations (49) and (50) take the form

$$\sum_n 1/h_n(\boldsymbol{\lambda})=(\bar{\rho}-\rho_l)\, N \tag{58}$$

$$\sum_n \left[n^{5/3}-(n-1)^{5/3}\right]/h_n(\boldsymbol{\lambda})=(y\,\bar{\rho}-\rho_l)\, N^{5/3}. \tag{59}$$

Combination of (58) and (59) leads to

$$(\bar{\rho}-\rho_l)\,\lambda_1+(y\,\bar{\rho}-\rho_l)\,\lambda_2=1 \tag{60}$$

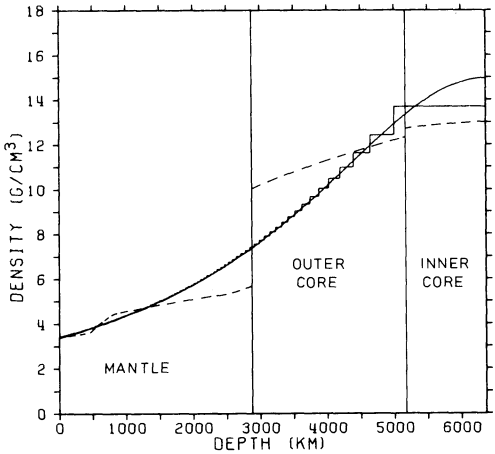

Fig. 1. Discrete ($N = 100$, step function) and continuous ($N \to \infty$) density distribution for lower density limit $\rho_l = 0$. Bullen's (1975) density distribution is represented by the dashed curve

which can be used to eliminate one of the unknowns from either Equation (58) or (59). The resulting nonlinear equation for the other unknown can be easily solved with standard methods.

For $N = 100$, $\rho_l = 0$ and $\rho_u = N\bar{\rho}$, for example,

$$\lambda_1 = 0.06667, \quad \lambda_2 = 0.13640 \quad [\mathrm{cm}^3/\mathrm{g}]. \tag{61}$$

The density $\bar{\rho}_n$ obtained from (57) by means of these parameters is shown in Figure 1 (step function).

With $\rho_u = N\bar{\rho}$ and $N \to \infty$, Equation (57) becomes

$$\bar{\rho}(r) = \rho_l + 1/h(r, \boldsymbol{\lambda}) \tag{62}$$

with

$$h(r, \boldsymbol{\lambda}) = \lambda_1 + \tfrac{5}{3}\lambda_2 (r/R)^2 \tag{63}$$

where λ_1 and λ_2 are solution of

$$\int_0^1 x^2\, dx/h(x\, R, \boldsymbol{\lambda}) = (\bar{\rho} - \rho_l)/3 \tag{64}$$

$$\int_0^1 x^4\, dx/h(x\, R, \boldsymbol{\lambda}) = (y\, \bar{\rho} - \rho_l)/5, \tag{65}$$

the continuous equivalent to Equations (58) and (59). The integrals can be evaluated in closed form and determination of the Lagrange multipliers proceeds in much the same way as in the discrete case discussed before. For $\rho_l = 0$ the following values are obtained.

$$\lambda_1 = 0.06674, \quad \lambda_2 = 0.13633 \quad [\mathrm{cm}^3/\mathrm{g}]. \tag{66}$$

The continuous density distribution computed with these Lagrange multipliers is shown in Figure 1 superimposed on the discrete density distribution. Both

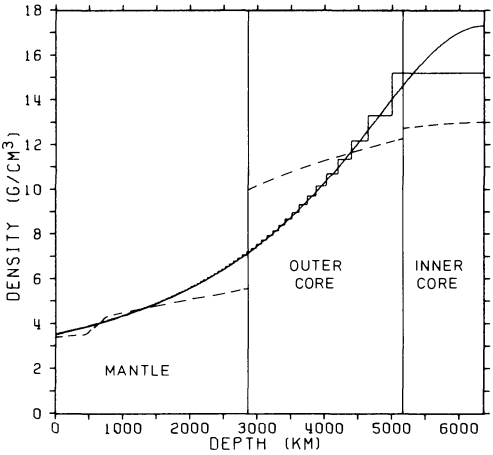

Fig. 2. Discrete ($N=100$, step function) and continuous ($N\to\infty$) density distribution for lower density limit $\rho_l=1$ g/cm^3. Bullen's (1975) density distribution is represented by the dashed curve

agree surprisingly well with the—according to Bullen (1975)—most likely density distribution indicated by dashed lines.

The density distribution changes when $\rho_l>0$ as shown in Figure 2 where $\rho_l=1$ g/cm^3 has been assumed. The Lagrange multipliers read

$$\lambda_1=0.06107, \quad \lambda_2=0.19924 \quad [\mathrm{cm}^3/\mathrm{g}] \tag{67}$$

$$\lambda_1=0.06120, \quad \lambda_2=0.19910 \quad [\mathrm{cm}^3/\mathrm{g}] \tag{68}$$

in the discrete ($N=100$) and continuous case, respectively. For depths greater than 4500 km the agreement between this curve and the "most likely" density distribution is poorer than in the previous case. This seems to be in disagreement with the fact that more information (i.e. that the density of the Earth is not less than 1 g/cm^3) has been supplied. In order to show that the uncertainty concerning the density has actually been reduced it is necessary to interpret the expectation values of the density in a different way.

Equations (54) and (55) give the probability distribution for the density in the n-th interval. The cumulative distribution function

$$\mathrm{cdf}_n(\rho)=\{1-\exp[(\rho_l-\rho)\,h_n(\boldsymbol{\lambda})]\}/\{1-\exp[(\rho_l-\rho_u)\,h_n(\boldsymbol{\lambda})]\} \tag{69}$$

gives the probability that $\rho_l\leqq\rho_n\leqq\rho$. For large $\rho_u\gg\rho_l$

$$\mathrm{cdf}_n(\rho)=1-\exp[-(\rho-\rho_l)/(\bar{\rho}_n-\rho_l)]\,. \tag{70}$$

Hence the probability that ρ_n be in the interval $[\rho_l,\bar{\rho}_n]$ is $1-1/e=0.632$ for any n; i.e. with a probability of 63.2% we expect the true density to be between the lower density limit and the expectation value of the density. This 63.2% interval is smaller in Figure 2 for depths less than ca. 4900 km.

From the form of the probability distribution (Eq. (54)) follows an important property of the average density

$$\rho=\frac{1}{N}\sum_n \rho_n\,. \tag{71}$$

The expectation value of this average density has been requested to be equal to the known mean value $\bar{\rho}$ of the Earth's density (Eq. (49)).

$$\frac{1}{N}\int p(\boldsymbol{\rho})\,(\sum_n \rho_n)\,dv=\frac{1}{N}\sum \bar{\rho}_n=\bar{\rho}.$$

It can be shown that the variance of the average density disappears for $N\to\infty$.

$$\begin{aligned}\operatorname{var}(\rho)&=\int p(\boldsymbol{\rho})\left(\frac{1}{N}\sum_n \rho_n-\bar{\rho}\right)^2 dv=\int p(\boldsymbol{\rho})\left(\frac{1}{N}\sum_n \rho_n\right)^2 dv-\bar{\rho}^2\\&=[\sum_{\substack{n,m\\n\neq m}} \bar{\rho}_n\,\bar{\rho}_m+2\sum_n \bar{\rho}_n^2]/N^2-\bar{\rho}^2=\sum_n \bar{\rho}_n^2/N^2.\end{aligned}\tag{72}$$

Since the $\bar{\rho}_n$ are bounded for $N\to\infty$ the last term in (72) vanishes for $N\to\infty$. Hence for the continuous density distribution the variance of the average density is equal to zero (weak law of large numbers).

An analoguous result holds for

$$\sum_n \rho_n\{[n/N]^{5/3}-[(n-1)/N]^{5/3}\}.$$

The expectation value of this expression is $y\,\bar{\rho}$ (see Eq. (50)) and its variance approaches zero for $N\to\infty$.

Concluding Remarks

This last example has not been presented to propose a new model for the density within the Earth. Rather it was to demonstrate that, based on the maximum entropy principle, a powerful method exists for extracting useful information from a very limited number of measured data.

For the density distribution within the Earth it leads to a quantification of qualitative statements (e.g. that the density increases with depth). Furthermore a probability (based on the given data and the prior prohability distribution) can be established that the density in a certain depth is within a certain density interval.

The maximum entropy concept is, of course, far from being exhaustively treated in this text. In fact many lines of thought have only been sketched, many problems have only been touched and require more detailed analysis. It is hoped that this paper provokes further investigation of the maximum entropy concept for handling of inverse problems. In any case it should be clear that this method must not be used indiscriminately and in no case substitute but only complement observations, measurements, and physical reasoning.

Acknowledgement. I am indebted to Deutsche Texaco Aktiengesellschaft for the permission to publish this paper.

References

Aczél, J., Daróczy, Z.: On measures of information and their characterizations. New York: Academic Press 1975

Akhiezer, N.I.: Theory of approximation. New York: Ungar 1956

Backus, G., Gilbert, F.: Numerical applications of a formalism for geophysical inverse problems. Geophys. J. **13**, 247-276, 1967

Backus, G., Gilbert, F.: The resolving power of gross Earth data. Geophys. J. **16**, 169-205, 1968

Backus, G., Gilbert, F.: Uniqueness in the inversion of inaccurate gross Earth data. Phil. Trans. Roy. Soc. London, Ser. A **266**, 123-192, 1970

Bullen, K.E.: The Earth's density. London: Chapman and Hall 1975

Cook, A.H.: The dynamical properties and internal constitutions of the Earth, the Moon, and the Planets. Quarterly J. Roy. astron. Soc. **12**, 154-168, 1971

Edward, J.A., Fitelson, M.M.: Notes on maximum entropy processing. IEEE Trans. Inf. Theory (Corresp.). IT-**19**, 232-234, 1973

Jaynes, E.T.: Prior probabilities. IEEE Trans. Systems Sci. Cybern. SSC-**4**, 227-241, 1968

Jensen, O. G., Ulrych, T.J.: An analysis of the perturbations on Barnard's star. Astron. J. **78**, 1104-1114, 1973

Kanasewich, E.R.: Time sequence analysis in geophysics. 2nd ed. Edmonton, Alberta: The University of Alberta Press 1975

Khinchin, R.T.: Mathematical foundations of information theory. New York: Dover Publications 1957

Papoulis, A.: A new class of Fourier series kernels. IEEE Trans. Circuit Theory CT-**20**, 101-107, 1973

Parker, R.L.: Inverse theory with grossly inadequate data. Geophys. J. **29**, 123-138, 1972

Phillips, J.D., Cox, A.: Spectral analysis of geomagnetic reversal time scales. Geophys. J. **45**, 19-33, 1976

Rao, C.R.: Linear statistical inference and its applications. New York: Wiley 1965

Rowlinson, I.S.: Probability, information and entropy. Nature **225**, 1196-1198, 1970

Smylie, D.E., Clarke, G.K.C., Ulrych, T.J.: Analysis of irregularities in the Earth's rotation. Methods in Computational Physics **13**, 391-430, 1973

Ulrych, T.J.: Maximum entropy power spectrum of long period geomagnetic reversals. Nature **235**, 218-219, 1972

Received September 15, 1976