Information **2011**, *2*, 277-301; doi:10.3390/info2020277

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

Spencer-Brown *vs.* Probability and Statistics: Entropy's Testimony on Subjective and Objective Randomness

Julio Michael Stern

Department of Applied Mathematics, Institute of Mathematics and Statistics, University of Sao Paulo, Rua do Matao 1010, Cidade Universitaria, 05508-090, Sao Paulo, Brazil;
E-Mail: jmstern@hotmail.com; Fax: +55-11-3819-3922

Received: 08 February 2011; in revised form: 22 March 2011 / Accepted: 23 March 2011 / Published: 4 April 2011

Abstract: This article analyzes the role of entropy in Bayesian statistics, focusing on its use as a tool for detection, recognition and validation of eigen-solutions. "Objects as eigen-solutions" is a key metaphor of the cognitive constructivism epistemological framework developed by the philosopher Heinz von Foerster. Special attention is given to some objections to the concepts of probability, statistics and randomization posed by George Spencer-Brown, a figure of great influence in the field of radical constructivism.

Keywords: Bayesian statistics; cognitive constructivism; eigen-solutions; maximum entropy; objective-subjective complementarity; randomization; subjective randomness

1. Introduction

In several already published articles, I defend the use of Bayesian Statistics in the epistemological framework of cognitive constructivism. In particular, I show how the FBST—The Full Bayesian Significance Test for precise hypotheses—can be used as a tool for detection, recognition and validation of eigen-solutions, see [1–12]. "Objects as eigen-solutions" is a key metaphor of cognitive constructivism as developed by the Austrian-American philosopher Heinz von Foerster, see [13]. For some recent applications in empirical science, see [14–20].

In Statistics, specially in the design of statistical experiments, Randomization plays a role which is in the very core of objective-subjective complementarity, a concept of great significance in the epistemological framework of cognitive constructivism as well as in the theory of Bayesian statistics.

The pivotal role of randomization in a well designed statistical experiment is that of a decoupling operation used to sever illegitimate functional links, thus avoiding spurious associations, breaking false influences, separating confounding variables, *etc.*, see [10] and [21].

The use of randomization in Statistics is an original idea of Charles Saunders Peirce and Joseph Jastrow, see [22,23]. Randomization is now a standard requirement for many scientific studies. In [8] and [10] I consider the position of C.S.Peirce as a forerunner of cognitive constructivism, based on the importance, relevance and coherence of his philosophical and scientific work. Among his several contributions, the introduction of randomization in statistical design stands indubitably out. In future articles, I hope to further expand the analysis of the role of Bayesian statistics in cognitive constructivism and provide other interesting applications.

I shall herein analyze some objections to the concepts of probability, statistics and randomization posed by George Spencer-Brown, a figure of great influence in the field of radical constructivism. Abstinence from statistical analysis and related quantitative methods may, at first glance, look like an idyllic fantasy island where many beautiful dreams come true. However, in my personal opinion, this position threatens to exile the cognitive constructivism epistemological framework to a limbo of powerless theories. In this article, entropy is presented as a cornerstone concept for the precise analysis and a key idea for the correct understanding of several important topics in probability and statistics. This understanding should help to clear the way for establishing Bayesian statistics as a preferred tool for scientific inference in mainstream cognitive constructivism.

In what follows, Section 2 corresponds to the first part of this article's title and elaborates upon "the case of Spencer-Brown *vs.* probability and statistics". Corresponding to the second part of the title, Section 3 provides "the testimony of entropy on subjective randomness". Section 4 gives "the testimony of entropy on objective randomness", presenting several mathematical definitions, theorems and algorithms. In this article, entropy based informational analysis is the key used to "solve" all the probability paradoxes and objections to statistical science posed by Spencer Brown. Section 4 is completely self-contained. Hence, a reader preferring to be exposed first to intuitions and motivations, can read the sections of this article in the order they are presented; meanwhile, a reader seeking a more axiomatic approach can start with Section 4. Section 5 presents our final conclusions.

2. Spencer-Brown, Probability and Statistics

In [24–26], Spencer-Brown analyzed some apparent paradoxes involving the concept of randomness, and concluded that the language of probability and statistics was inappropriate for the practice of scientific inference. In subsequent work, [27], he reformulates classical logic using only a generalized *nor* operator (marked *not-or*, unmarked *or*), that he represents à la mode of Charles Saunders Peirce or John Venn, by a graphical boundary or distinction mark, see [28–34].

Making (or arbitrating) distinctions is, according to Spencer-Brown, the basic (if not the only) operation of human knowledge, an idea that has either influenced or been directly explored by several authors in the radical constructivist movement. The following quotations, from [26] p. 23, p. 66 and p. 105, are typical arguments used by Spencer-Brown in his rejection of probability and statistics:

> *Retroactive reclassification of observations in one of the scientist's most important tools, and we shall meet it again when we consider statistical arguments. (p. 23)*
>
> *We have found so far that the concept of probability used in statistical science is meaningless in its own terms; but we have found also that, however meaningful it might have been, its meaningfulness would nevertheless have remained fruitless because of the impossibility of gaining information from experimental results, however significant. This final paradox, in some ways the most beautiful, I shall call the Experimental Paradox (p. 66).*
>
> *The essence of randomness has been taken to be absence of pattern. But what has not hitherto been faced is that the absence of one pattern logically demands the presence of another. It is a mathematical contradiction to say that a series has no pattern; the most we can say is that it has no pattern that anyone is likely to look for. The concept of randomness bears meaning only in relation to the observer: If two observers habitually look for different kinds of pattern they are bound to disagree upon the series which they call random (p. 105).*

Several authors concur, at least in part, with my opinion about Spencer-Brown's technical analysis of probability and statistics, see [35–39]. In Section 3, I carefully explain why I disagree with it. In some of my arguments, which are are based on information theory and the notion of entropy, I dissent from Spencer-Brown's interpretation of measures of order-disorder in sequential signals. In [40–44], some of the basic concepts in this area are reviewed with a minimum of mathematics. For more advanced developments see [45–47].

I also disapprove some of Spencer Brown's proposed methodologies to detect "relevant" event sequences, that is, his criteria to "mark distinct patterns" in empirical observations. My objections have a lot in common with the standard caveats against *ex post facto* "fishing expeditions" for interesting outcomes, or simple *post hoc* "sub-group analysis" in experimental data banks. This kind of retroactive or retrospective data analyses is considered a questionable statistical practice, and pointed as the culprit of many misconceived studies, misleading arguments and mistaken conclusions. The literature on statistical methodology for clinical trials has been particularly keen in warning against this kind of practice. See [48,49] for two interesting papers addressing this specific issue and published in high impact medicine journals less than a year before I wrote this text. When consulting for pharmaceutical companies or advising in the design of statistical experiments, I often find it useful to quote Conan Doyle's Sherlock Holmes, in The Adventure of Wisteria Lodge:

> *Still, it is an error to argue in front of your data. You find yourself insensibly twisting them around to fit your theories.*

Finally, I am also suspicious or skeptical about the intention behind some applications of Spencer-Brown's research program, including the use of extrasensory empathic perception for coded message communication, exercises on object manipulation using paranormal powers, *etc*. Unable to reconcile his psychic research program with statistical science, Spencer-Brown had no regrets in disqualifying the later, as he clearly stated in the prestigious scientific journal *Nature*, see pp. 594–595 of [25]:

> *[On telepathy:] Taking the psychical research data (that is, the residuum when fraud and incompetence are excluded), I tried to show that these now threw more doubt upon existing pre-suppositions in the theory of probability than in the theory of communication.*
>
> *[On psychokinesis:] If such an 'agency' could thus 'upset' a process of randomizing, then all our conclusions drawn through the statistical tests of significance would be equally affected, including the conclusions about the 'psychokinesis' experiments themselves. (How are the target numbers for the die throws to be randomly chosen? By more die throws?) To speak of an 'agency' which can 'upset' any process of randomization in an uncontrollable manner is logically equivalent to speaking of an inadequacy in the theoretical model for empirical randomness, like the luminiferous ether of an earlier controversy, becomes, with the obsolescence of the calculus in which it occurs, a superfluous term.*

Spencer-Brown's conclusions in [24–26], including his analysis of probability, were considered to be controversial (if not unreasonable or extravagant) even by his own colleagues at the Society of Psychical Research, see [50,51]. It seems that current research in this area, even not being free (or afraid) of criticism, has abandoned the path of naïve confrontation with statistical science, see [52,53]. For additional comments, see [54–57].

Curiously, Charles Saunders Peirce and his student Joseph Jastrow, who introduced the idea of randomization in statistical trials, also struggled with some of the very same dilemmas faced by Spencer-Brown, namely, the eventual detection of distinct patterns or seemingly ordered (sub)strings in a long random sequence. Peirce and Jastrow did not have at their disposal the heavy mathematical artillery I have quoted in the previous paragraphs. Nevertheless, as experienced explorers that are not easily lured, when traveling in desert sands, by the mirage of a misplaced oasis, these intrepid pioneers were able to avoid the conceptual pitfalls that lead Spencer-Brown so far astray. For more details see [10], [22,23] and [58–60].

As stated in the introduction, the cognitive constructivist framework can be supported by the FBST, a non-decision theoretic formalism drawn from Bayesian statistics, see [1] and [3–5]. The FBST was conceived as a tool for validating objective knowledge of eigen-solutions and, as such, can be easily integrated to the epistemological framework of cognitive constructivism in scientific research practice. Contrasting our distinct views of cognitive constructivism, it is not at all surprising that I have come to conclusions concerning the use of probability and statistics, and also to the relation between probability and logic, that are fundamentally different from those of Spencer-Brown.

3. Pseudo, Quasi and Subjective Randomness

The focus of the present section are the properties of "natural" and "artificial" random sequences. The implementation of probabilistic algorithms require good random number generators, (RNGs). These algorithms include: Numerical integration methods such as Monte Carlo or Markov Chain Monte Carlo (MCMC); evolutionary computing and stochastic optimization methods such as genetic programming and simulated annealing; and also, of course, the efficient implementation of randomization methods.

The most basic random number generator replicates i.i.d. (independent and identically distributed) random variables uniformly distributed in the unit interval, $[0, 1]$. From this basic uniform generator one

gets a uniform generator in the d-dimensional unit box, $[0,1]^d$, and, from the later, non-linear generators for many other multivariate distributions, see [61,62].

Historically, the technology of random number generators was developed in the context of Monte Carlo methods. The nature of Monte Carlo algorithms makes them very sensitive to correlations, auto-correlations and other statistical properties of the random number generator used in its implementation. Hence, in this context, the statistical properties of "natural" and "artificial" random sequences came to close scrutiny. For the aforementioned historical and technological reasons, Monte Carlo methods are frequently used as a benchmark for testing the properties of these generators. Hence, although Monte Carlo methods proper lie outside the scope of this article, we shall keep them as a standard application benchmark in our discussions.

The clever ideas and also the caveats of engineering good random number generators are in the core of many paradoxes found by Spencer-Brown. The objective of this section is to explain the basic ideas behind these generators and, in so doing, avoid the conceptual traps and pitfalls that took Spencer-Brown analyses so much off course.

3.1. Random and Pseudo-Random Number Generators

The concept of randomness is usually applied to a variable or a process (to be generated or observed) involving some uncertainty. The following definition is presented at p. 10 of [61]:

> *A random event is an event which has a chance of happening, and probability is a numerical measure of that chance.*

Monte Carlo, and several other probabilistic algorithms, require a random number generator. With the last definition in mind, engineering devices based on sophisticated physical processes have been built in the hope of offering a source of "true" random numbers. However, these special devices were cumbersome, expensive, not portable nor easily available, and often unreliable. Moreover, practitioners soon realized that simple deterministic sequences could successfully be used to emulate a random generator, as stated in the following quotes (our emphasis) at p. 26 of [61] and p. 15 of [62]:

> *For electronic digital computers it is most convenient to calculate a sequence of numbers one at a time as required, by a completely specified rule which is, however, so devised that no* **reasonable** *statistical test will detect any significant departure from randomness. Such a sequence is called* pseudorandom. *The great advantage of a specified rule is that the sequence can be exactly reproduced for purposes of computational checking.*

> *A sequence of* pseudorandom *numbers (U_i) is a deterministic sequence of numbers in $[0,1]$ having the same* **relevant** *statistical properties as a sequence of random numbers.*

Many deterministic random emulators used today are Linear Congruential Pseudo-Random Generators (LCPRG), as in the following example:

$$x_{i+1} = (ax_i + c) \bmod m$$

where the multiplier a, the increment c and the modulus m should obey the conditions: (i) c and m are relatively prime; (ii) $a-1$ is divisible by all prime factors of m; (iii) $a-1$ is a multiple of 4 if m is

a multiple of 4. LCPRG's are fast and easy to implement if m is taken as the computer's word range, 2^s, where s is the computer's word size, typically $s = 32$ or $s = 64$. The LCPRG's starting point, x_0, is called the seed. Given the same seed the LCPG will reproduce the same sequence, a very convenient feature for tracing, debugging and verifying application programs.

However, LCPRG's are not an universal solution. For example, it is trivial to devise some statistics whose behaviour will be far from random, see [63]. There the importance of the words **reasonable** and **relevant** in the last quotations becomes clear: For most practical applications these statistics are irrelevant. LCPRG's can also exhibit very long range auto-correlations and, unfortunately, these are more likely to affect long simulated time series required in some special applications. The composition of several LCPRG's by periodic seed refresh may mitigate some of these difficulties, see [62]. LCPRG's are also not appropriate to some special applications in cryptography, see [64]. Current state of the art generators are given in [65,66].

3.2. Chance is Lumpy—Quasi-Random Generators

"*Chance is Lumpy*" is Robert Abelson's First Law of Statistics, stated in p. XV of [67]. The probabilistic expectation is a linear operator, that is, $E(Ax + b) = AE(x) + b$, where x in random vector and A and b are a determined matrix and vector. The Covariance operator is defined as $\text{Cov}(x) = E((x - E(x)) \otimes (x - E(x)))$. Hence, $\text{Cov}(Ax + b) = A\text{Cov}(x)A'$. Therefore, given n i.i.d. scalar variables, $x_i \,|\, \text{Var}(x_i) = \sigma^2$, the variance of their mean, $m = (1/n)\mathbf{1}'x$ (notice the simplified vector notation $\mathbf{1} = [1, 1 \ldots, 1]$), is given by

$$\frac{1}{n}\mathbf{1}' \,\text{diag}(\sigma^2\mathbf{1})\, \frac{1}{n}\mathbf{1} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix} \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix} = \sigma^2/n$$

Hence, mean values of iid random variables converge to their expected values at a rate of $1/\sqrt(n)$.

Quasi-random sequences are deterministic sequences built not to emulate random sequences, as pseudo-random sequences do, but to achieve faster convergence rates. For d-dimensional quasi-random sequences, an appropriate measure of fluctuation, called discrepancy, only grows at a rate of $\log(n)^d$, hence growing much slower than $\sqrt(n)$. Therefore, the convergence rate corresponding to quasi-random sequences, $\log(n)^d/n$, is much faster than the one corresponding to (pseudo) random sequences, $\sqrt(n)/n$. Figure 1 allows the visual comparison of typical (pseudo) random (left) and quasi-random (right) sequences in $[0, 1]^2$. By visual inspection we see that the points of the quasi-random sequence are more "homogeneously scattered", that is, they do not "clump together", as the point of the (pseudo) random sequence often do.

Let us consider an axis-parallel rectangles in the unit box,

$$R = [a_1, b_1] \times [a_2, b_2] \times \ldots [a_d, b_d] \subseteq [0, 1]^d$$

The discrepancy of the sequence $s_{1:n}$ in box R, and the overall discrepancy of the sequence are defined as

$$D(s_{1:n}, R) = n\text{Vol}(R) - |s_{1:n} \cap R| \;, \quad D(s_{1:n}) = \sup_{R \subseteq [0,1[^d} |D(s_{1:n}, R)|$$

It is possible to prove that the discrepancy of the Halton-Hammersley sequence, defined next, is of order $O(\log(n)^{d-1})$, see chapter 2 of [68].

Halton-Hammersley sets: Given $d-1$ distinct prime numbers, $p(1), p(2), \ldots p(d-1)$, the i-th point, x^i, in the Halton-Hammersley set, $\{x^1, x^2, \ldots x^n\}$, is

$$x^i = \left[i/n, r_{p(1)}(i), r_{p(2)}(i), \ldots r_{p(d-1)}(i)\right]' \ , \ \text{for} \ \ i = 1 : n-1 \ , \ \ \text{where,}$$

$$i = a_0 + p(k)a_1 + p(k)^2 a_2 + p(k)^3 a_3 + \ldots \ , \ \ r_{p(k)}(1) = \frac{a_0}{p(k)} + \frac{a_1}{p(k)^2} + \frac{a_2}{p(k)^3} + \ldots$$

That is, the $(k+1)$-th coordinate of x^i, $x^i_{k+1} = r_{p(k)}(i)$, is obtained by the bit (or digit) reversal of i written in $p(k)$-adic or base $p(k)$ notation.

The Halton-Hammersley set is a generalization of van der Corput set, built in the bidimensional unit square, $d = 2$, using the first prime number, $p = 2$. The following example, from p. 33 of [61] and p. 117 of [69], builds the 8-point van der Corput set, expressed in binary and decimal notation.

```
function x= corput(n,b)
% size n base b v.d.corput set
m=floor(log(n)/log(b));
u=1:n;  D=[];
for i=0:m
  d= rem(u,b);
  u= (u-d)/b;
  D= [D; d];
end
x=((1./b').^(1:(m\ma1)))*D;
```

Decimal		Binary	
i	$r_2(i)$	i	$r_2(i)$
1	0.5	1	0.1
2	0.21	10	0.01
3	0.75	11	0.11
4	0.125	100	0.001
5	0.625	101	0.101
6	0.375	110	0.011
7	0.875	111	0.111
8	0.0625	1000	0.0001

Quasi-random sequences, also known as low-discrepancy sequences, can substitute pseudo-random sequences in some applications of Monte Carlo methods, achieving higher accuracy with less computational effort, see [70–72]. Nevertheless, since by design the points of a quasi-random sequence tend to avoid each other, strong (negative) correlations are expected to appear. In this way, the very reason that can make quasi-random sequences so helpful, can ultimately impose some limits to their applicability. Some of these problems are commented in p. 766 of [73]:

> *First, quasi-Monte Carlo methods are valid for integration problems, but may not be directly applicable to simulations, due to the correlations between the points of a quasi-random sequence. ... A second limitation: the improved accuracy of quasi-Monte Carlo methods is generally lost for problems of high dimension or problems in which the integrand is not smooth.*

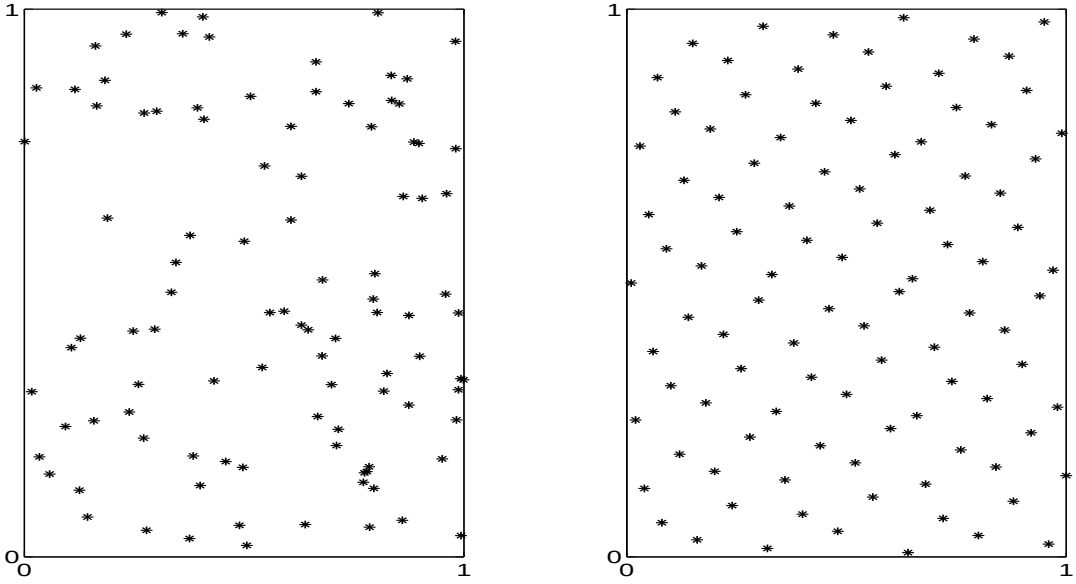
3.3. Subjective Randomness and Its Paradoxes

When asked to look at patterns like those in Figure 1, many subjects perceive the quasi-random set as "more random" than the (pseudo) random set. How can this paradox be explained? This was the topic of

many psychological studies in the field of subjective randomness. The quotation in the next paragraph is from one of these studies, p. 306 in [36], emphasis are ours:

> *One major source of confusion is the fact that randomness involves two distinct ideas:* **process** *and* **pattern***, [74]. It is natural to think of randomness as a process that generates unpredictable outcomes (stochastic process according to [75]). Randomness of a* **process** *refers to the* **unpredictability** *of the individual event in the series [76,77]. This is what Spencer Brown [26] calls* **primary randomness***. However, one usually determines the randomness of the process by means of its output, which is supposed to be* **patternless***. This kind of randomness refers, by definition, to a sequence. It is labeled* **secondary randomness** *by Spencer Brown. It requires that all symbol types, as well as all ordered pairs (diagrams), ordered triplets (trigrams)... n-grams in the sequence be equiprobable. This definition could be valid for any n only in infinite sequences, and it may be approximated in finite sequences only up to ns much smaller than the sequence's length. The entropy measure of randomness is based on this definition, see chapter 1 and 2 of [41].*
>
> *These two aspects of randomness are closely related. We ordinarily expect outcomes generated by a random process to be patternless. Most of them are. Conversely, a sequence whose order is random supports the hypothesis that it was generated by a random mechanism, whereas sequences whose order is not random cast doubt on the random nature of the generating process.*

Figure 1. (Pseudo)—random and quasi-random point sets on the unit box.



Spencer-Brown was intrigued by the apparent incompatibility of the notions of primary and secondary randomness. The apparent collision of these two notions generates several interesting paradoxes, taking Spencer-Brown to question the applicability of the concept of randomness in particular and probability and statistical analysis in general, see [24–26], and also [35], [38,39], [54–57] and [78], In fact, several subsequent psychological studies were able to confirm that, for many subjects, the intuitive

or common-sense perception of primary and secondary randomness are quite discrepant. However, a careful mathematical analysis makes it possible to reconcile the two notions of randomness. These are the topics discussed in this section.

The relation between the joint and conditional entropy for a pair of random variables, see Section 4,

$$H(i,j) = H(j) + H(i\,|\,j) = H(i) + H(j\,|\,i)$$

motivates the definition of first, second and higher order entropies, defined over the distribution of words of size m in a string of letters from an alphabet of size a.

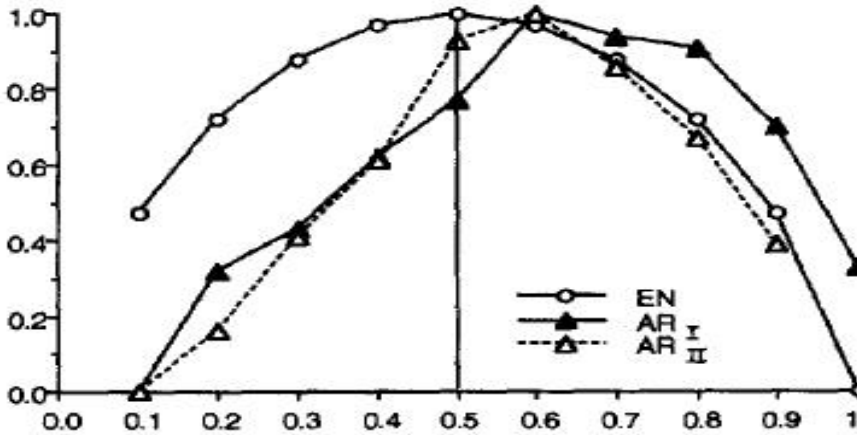
$$H_1 = \sum_j p(j) \log p(j)\ , \quad H_2 = \sum_{i,j} p(i)p(j\,|\,i) \log p(j\,|\,i)$$

$$H_3 = \sum_{i,j,k} p(i)p(j\,|\,i)p(k\,|\,i,j) \log p(k\,|\,i,j)\ \ldots$$

It is possible to use these entropy measures to assess the disorder or lack of pattern in a given finite sequence, using the empirical probability distributions of single letters, pairs, triplets, *etc.* However, in order to have a significant empirical distribution of m-plets, any possible m-plet must be well represented in the sequence, that is, the word size, m, is required to be very short relative to the sequence log-size, that is, $m << \log_a(n)$.

In the article [36], Figure 2 displays the typical perceived or apparent randomness of Boolean (0-1) bit sequences, represented as black-and-white pixel in linear arrays, *versus* the second order entropy of the same strings, see also [41]. Clearly, there is a remarkable bias of the apparent randomness relative to the entropic measure.

Figure 2. EN, H_2-entropy *vs.* AR, apparent randomness. Probability of black-white pixel alternation.



This effect is known as the *gambler's fallacy* when betting on *cool spots*. It consists of expecting the random sequence to "compensate" finite average fluctuations from expected values. This effect is also described in p. 303 of [36]:

> *When people invent superfluous explanations because they perceive patterns in random phenomena, they commit what is known in statistical parlance as Type I error. The other*

> *way of going awry, known as Type II error, occurs when one dismisses stimuli showing some regularity as random. The numerous randomization studies in which participants generated too many alternations and viewed this output as random, as well as the judgments of overalternating sets as maximally random in the perception studies, were all instances of type II error in research results.*

It is known that other gamblers exhibit the opposite behavior, preferring to bet on *hot spots*, expecting the same fluctuations to occur repeatedly. These effects are the consequence of a perceived coupling, by a negative or positive correlation or other measure of association, between non overlapping segments that are in fact supposed to be decoupled, uncorrelated or have no association, that is, to be independent. For a statistical analysis, see [58,59]. A possible psychological explanation of the gambler's fallacy is given by the constructivist theory of Jean Piaget, see [79], as quoted in p. 316 of [36], in which any "lump" in the sequence is (miss) perceived as non-random order:

> *In analogy to Piaget's operations, which are conceived as internalized actions, perceived randomness might emerge from hypothetical action, that is, from a thought experiment in which one describes, predicts, or abbreviates the sequence. The harder the task in such a thought experiment, the more random the sequence is judged to be.*

The same hierarchical decomposition scheme used for higher order conditional entropy measures can be adapted to measure the disorder or patternless of a sequence, relative to a given subject's model of "computer" or generation mechanism. In the case of a discrete string, this generation model could be, for example, a deterministic or probabilistic Turing machine, a fixed or variable length Markov chain, *etc.* It is assumed that the model is regulated by a code, program or vector parameter, θ, and outputs a data vector or observed string, x. The hierarchical complexity measure of such a model emulates the Bayesian prior and conditional likelihood decomposition, $H(p(\theta, x)) = H(p(\theta)) + H(p(x \mid \theta))$, that is, the total complexity is given by the complexity of the program plus the complexity of the output given the program. This is the starting point for several complexity models, like Andrey Kolmogorov, Ray Solomonoff and Gregory Chaitin's computational complexity models, Jorma Rissanen's Minimum Description Length (MDL), and Chris Wallace and David Boulton's Minimum Message Length (MML). All these alternative complexity models can also be used to successfully reconcile the notions of primary and secondary randomness, showing that they are asymptotically equivalent, see [80–85].

4. Entropy and Its Use in Mathematical Statistics

Entropy is the cornerstone concept of the preceding section, used as a central idea in the understanding of order and disorder in stochastic processes. Entropy is the key that allowed us to unlock the mysteries and solve the paradoxes of subjective randomness, making it possible to reconcile the notions of unpredictability of stochastic process and patternless of randomly generated sequences. Similar entropy based arguments reappear, in more abstract, subtle or intricate forms, in the analysis of technical aspects of Bayesian statistics like, for example, the use of prior and posterior distributions and the interpretation of their informational content. This section gives a short review covering the definition of entropy, its main properties, and some of its most important uses in mathematical statistics.

The origins of the entropy concept lay in the fields of Thermodynamics and Statistical Physics, but its applications have extended far and wide to many other phenomena, physical or not. The entropy of a probability distribution, $H(p(x))$, is a measure of uncertainty (or impurity, confusion) in a system whose states, $x \in \mathcal{X}$, have $p(x)$ as probability distribution. We follow closely the presentation in the following references. For the basic concepts, see [42] and [86–89]. For maximum entropy (MaxEnt) characterizations, see [45] and [90]. For numerical optimization methods for MaxEnt problems, see [91–95]. For posterior asymptotic convergence, see [96]. For a detailed analysis of the connection between MaxEnt optimization and Bayesian statistics' formalisms, that is, for a deeper view of the relation between MaxEnt and Bayes' rule updates, see [97].

4.1. Convexity

This section introduces the notion of convexity, a concept at the heart of the definition of entropy and generalized directed divergences. Convexity arguments are also needed to prove, in the following sections, important properties of entropy and its generalizations. In this section we use the following notations: $\mathbf{0}$ and $\mathbf{1}$ are the origin and unit vector of appropriate dimension. Subscripts are used as an element index in a vector or as a row index in a matrix, and superscripts are used as an index for distinct vectors or as a column index in a matrix.

Definition: A region $S \in R^n$ is Convex iff, for any two points, $x^1, x^2 \in S$, and weights $0 \leq l_1, l_2 \leq 1 \,|\, l_1 + l_2 = 1$, the convex combination of these two points remains in S, *i.e.* $l_1 x^1 + l_2 x^2 \in S$.

Theorem: Finite Convex Combination: A region $S \in R^n$ is Convex iff any (finite) convex combination of its points remains in the region, *i.e.*, $\forall\ \mathbf{0} \leq l \leq \mathbf{1} \,|\, \mathbf{1}'l = 1$, $X = [x^1, x^2, \ldots x^m]$, $x^j \in S$,

$$X\,l = \begin{bmatrix} x_1^1 & x_1^2 & \ldots & x_1^m \\ x_2^1 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ \ldots \\ l_m \end{bmatrix} \in S$$

Proof: By induction in the number of points, m.

Definition: The Epigraph of the function $\varphi : R^n \to R$ is the region of X "above the graph" of φ, *i.e.*,

$$\text{Epi}\,(\varphi) = \left\{ x \in R^{n+1} \,|\, x_{n+1} \geq \varphi\left([x_1, x_2, \ldots, x_n]' \right) \right\}$$

Definition: A function φ is convex iff its epigraph is convex. A function φ is concave iff $-\varphi$ is convex.

Theorem: A differentiable function, $\varphi : R \to R$, with non negative second derivative is convex.

Proof: Consider $x^0 = l_1 x^1 + l_2 x^2$, and the Taylor expansion around x^0,

$$\varphi(x) = \varphi(x^0) + \varphi'(x^0)(x - x^0) + (1/2)\varphi''(x^*)(x - x^0)^2$$

where x^* is an appropriate intermediate point. If $\varphi''(x^*) > 0$ the last term is positive. Now, making $x = x^1$ and $x = x^2$ we have, respectively, that $\varphi(x^1) \geq \varphi(x^0) + \varphi'(x^0) l_1 (x^1 - x^2)$

and $\varphi(x^2) \geq \varphi(x^0) + \varphi'(x^0)l_2(x^2 - x^1)$ multipying the first inequality by l_1, the second by l_2, and adding them, we obtain the desired result.

Theorem: Jensen Inequality: If φ is a convex function,

$$\mathrm{E}\left(\varphi(x)\right) \geq \varphi\left(\mathrm{E}(X)\right)$$

For discrete distributions the Jensen inequality is a special case of the finite convex combination theorem. Arguments of Analysis allow us to extend the result to continuous distributions.

4.2. Boltzmann-Gibbs-Shannon Entropy

If $H(p(x))$ is to be a measure of uncertainty, it is reasonable that it should satisfy the following list of requirements. For the sake of simplicity, we present several aspects of the theory in finite spaces.

(1) If the system has n possible states, $x_1, \ldots x_n$, the entropy of the system with a given distribution, $p_i \equiv p(x_i)$, is a function

$$H = H_n(p_1, \ldots, p_n)$$

(2) H is a continuous function.

(3) H is a function symmetric in its arguments.

(4) The entropy is unchanged if an impossible state is added to the system, *i.e.*,

$$H_n(p_1, \ldots p_n) = H_{n+1}(p_1, \ldots p_n, 0)$$

(5) The system's entropy is minimal and null when the system is fully determined, *i.e.*,

$$H_n(0, \ldots, 0, 1, 0, \ldots 0) = 0$$

(6) The system's entropy is maximal when all states are equally probable, *i.e.*,

$$\{\, \frac{1}{n}\mathbf{1} \,\} = \arg\max H_n$$

(7) A system maximal entropy increases with the number of states, *i.e.*,

$$H_{n+1}\left(\frac{1}{n+1}\mathbf{1}\right) > H_n\left(\frac{1}{n}\mathbf{1}\right)$$

(8) Entropy is an extensive quantity, *i.e.*, given two independent systems, with distributions p and q, the entropy of the composite system is additive, *i.e.*,

$$H_{nm}(r) = H_n(p) + H_m(q)\ , \quad r_{i,j} = p_i\, q_j$$

The Boltzmann-Gibbs-Shannon measure of entropy,

$$H_n(p) = -I_n(p) = -\sum_{i=1}^{n} p_i \log(p_i) = -\,\mathrm{E}_i \log(p_i)\ , \qquad 0\,\log(0) \equiv 0$$

satisfies requirements (1) to (8), and is the most usual measure of entropy. In Physics it is usual to take the logarithm in Napier base, while in Computer Science it is usual to take base 2 and in Engineering it

is usual to take base 10. The opposite of the entropy, $I(p) = -H(p)$, the Negentropy, is a measure of Information available about the system.

For the Boltzmann-Gibbs-Shannon entropy we can extend requirement 8, and compute the composite Negentopy even without independence:

$$\begin{aligned} I_{nm}(r) &= \sum_{i=1,j=1}^{n,m} r_{i,j}\ \log(r_{i,j}) = \sum_{i=1,j=1}^{n,m} p_i \mathrm{Pr}(j\,|\,i)\ \log\left(p_i \mathrm{Pr}(j\,|\,i)\right) \\ &= \sum_{i=1}^{n} p_i \log(p_i) \sum_{j=1}^{m} \mathrm{Pr}(j\,|\,i) + \sum_{i=1}^{n} p_i \sum_{j=1}^{m} \mathrm{Pr}(j\,|\,i)\ \log\left(\mathrm{Pr}(j\,|\,i)\right) \\ &= I_n(p) + \sum_{i=1}^{n} p_i\, I_m(q^i) \quad \text{where}\ ,\ \ q^i_j = \mathrm{Pr}(j\,|\,i) \end{aligned}$$

If we add this last identity as item number 9 in the list of requirements, we have a characterization of Boltzmann-Gibbs-Shannon entropy, see [87–89].

Like many important concepts, this measure of entropy was discovered and re-discovered several times in different contexts, and sometimes the uniqueness and identity of the concept was not immediately recognized. A well known anecdote refers the answer given by von Neumann, after Shannon asked him how to call a "newly" discovered concept in Information Theory. As reported by Shannon in p. 180 of [98]:

> *"My greatest concern was what to call it. I thought of calling it information, but the word was overly used, so I decided to call it uncertainty. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."*

4.3. Csiszar's Divergence

In order to check that requirement (6) is satisfied, we can use (with $q \propto 1$) the following lemma:

Lemma: Shannon Inequality.

If p and q are two distributions over a system with n possible states, and $q_i \neq 0$, then the Information of p Relative to q, $I_n(p,q)$, is positive, except if $p = q$, when it is null,

$$I_n(p,q) \equiv \sum_{i=1}^{n} p_i\ \log\left(\frac{p_i}{q_i}\right), \quad I_n(p,q) \geq 0\ \ ,\ I_n(p,q) = 0 \Rightarrow p = q$$

Proof: By Jensen inequality, if φ is a convex function,

$$\mathrm{E}\left(\varphi(x)\right) \geq \varphi\left(\mathrm{E}(X)\right)$$

Taking

$$\begin{aligned} \varphi(t) &= t\ln(t) \ \text{ and } \ t_i = \frac{p_i}{q_i} \\ \mathrm{E}_q\left(t\right) &= \sum_{i=1}^{n} q_i\, \frac{p_i}{q_i} = 1 \\ I_n(p,q) &= \sum q_i\ t_i \log t_i \geq 1\ \log(1) = 0 \end{aligned}$$

Shannon's inequality motivates the use of the Relative Information as a measure of (non symmetric) "distance" between distributions. In Statistics this measure is known as the Kullback-Leibler distance. The denominations Directed Divergence or Cross Information are used in Engineering. The proof of Shannon inequality motivates the following generalization of divergence:

Definition: Csiszar's φ-divergence.

Given a convex function φ,

$$\begin{aligned} d_\varphi(p,q) &= \sum_{i=1}^n q_i \, \varphi\left(\frac{p_i}{q_i}\right) \\ 0\,\varphi\left(\frac{0}{0}\right) &= 0 \;,\; 0\,\varphi\left(\frac{c}{0}\right) = c \lim_{t\to\infty} \frac{\varphi(t)}{t} \end{aligned}$$

For example, we can define the quadratic and the absolute divergence as

$$\begin{aligned} \chi^2(p,q) &= \sum \frac{(p_i - q_i)^2}{q_i} \;, \quad \text{for } \varphi(t) = (t-1)^2 \\ Ab(p,q) &= \sum \frac{|p_i - q_i|}{q_i} \;, \quad \text{for } \varphi(t) = |t-1| \end{aligned}$$

4.4. Maximum Entropy under Constraints

This section analyzes solution techniques for some problems formulated as entropy maximization. The results obtained in this section are needed to obtain some fundamental principles of Bayesian statistics, presented in the following sections. This section also presents the Bregman algorithm for solving constrained maxent problems on finite distributions. The analysis of small problems (far from asymptotic conditions) poses many interesting questions in the study of subjective randomness, an area so far neglected in the literature.

Given a prior distribution, q, we would like to find a vector p that minimizes the Relative Information $I_n(p,q)$, where p is under the constraint of being a probability distribution, and maybe also under additional constraints over the expectation of functions taking values on the system's states, that is, we want

$$\{p^*\} = \arg\min I_n(p,q) \;,\; p \geq 0 \;\mid \mathbf{1}'p = 1 \text{ and } Ap = b \;,\; A\;(m-1)\times n$$

p^* is the Minimum Information or Maximum Entropy (MaxEnt) distribution, relative to q, given the constraints $\{A, b\}$. We can write the probability normalization constraint as a generic linear constraint, including $\mathbf{1}$ and 1 as the m-th (or 0-th) rows of matrix A and vector b. So doing, we do not need to keep any distinction between the normalization and the other constraints. In this article, the operators $\odot$ and $\oslash$ indicate the point (element) wise product and division between matrices of same dimension.

The Lagrangian function of this optimization problem, and its derivatives are:

$$L(p,w) = p' \log(p \oslash q) + w'(b - Ap)$$

$$\frac{\partial L}{\partial p_i} = \log(p_i/q_i) + 1 - w'A^i \;,\; \frac{\partial L}{\partial w_k} = b_k - A_k p$$

Equating the $n+m$ derivatives to zero, we have a system with $n+m$ unknowns and equations, giving viability and optimality conditions (VOCs) for the problem:

$$p_i = q_i \exp\left(w' A^i - 1\right) \text{ or } p = q \odot \exp\left((w'\, A)' - \mathbf{1}\right)$$
$$A_k p = b_k,\ p \geq 0$$

We can further replace the unknown probabilities, p_i, writing the VOCs only on w, the dual variables (Lagrange multipliers),

$$h_k(w) \equiv A_k \left(q \odot \exp\left((w'\, A)' - \mathbf{1}\right)\right) - b_k = 0$$

The last form of the VOCs motivates the use of iterative algorithms of Gauss-Seidel type, solving the problem by cyclic iteration. In this type of algorithm, one cyclically "fits" one equation of the system, for the current value of the other variables. For a detailed analysis of this type of algorithm, see [91–95] and [99].

Bregman Algorithm:

Initialization: Take $t = 0$, $w^t \in R^m$, and

$$p_i^t = q_i \exp\left(w^{t\prime} A^i - 1\right)$$

Iteration step: for $t = 1, 2, \ldots$, Take

$$k = (t \bmod m) \quad \text{and} \quad \nu \,|\, \varphi(\nu) = 0\,, \quad \text{where}$$

$$\begin{aligned} w^{t+1} &= \left[w_1^t, \ldots w_{k-1}^t, w_k^t + \nu, w_{k+1}^t, \ldots w_m^t\right]' \\ p_i^{t+1} &= q_i \exp(w^{t+1\prime} A^i - 1) = p_i^t \exp(\nu A_k^i) \\ \varphi(\nu) &= A_k p^{t+1} - b_k \end{aligned}$$

From our discussion of Entropy optimization under linear constraints, it should be clear that the maximum relative entropy distribution for a system under constraints on the expectation of functions taking values on the system's states,

$$E_{p(x)} a_k(x) = \int a_k(x) p(x) dx = b_k$$

(including the normalization constraint, $a_0 = \mathbf{1}, b_0 = 1$) has the form

$$p(x) = q(x) \exp\left(-\theta_0 - \theta_1\, a_1(x) - \theta_2\, a_2(x) \ldots\right)$$

Notice that we took $\theta_0 = -(w_0 - 1)$, $\theta_k = -w_k$, and we have also indexed the state i by variable x, so to write the last equation in the standard form used in the statistical literature.

Several distributions commonly used in Statistics can be interpreted as MaxEnt densities (relative to the uniform distribution, if not otherwise stated) given some constraints over the expected value of state functions. For example:

The Normal distribution:

$$f(x \,|\, n, \beta, R) = c(R) \exp(\, -\frac{n}{2}(x - \beta)' R (x - \beta)\,)$$

is characterized as the distribution of maximum entropy on R^n, given the expected values of its first and second moments, *i.e.*, mean vector β and inverse covariance or precision matrix R.

The Wishart distribution:

$$f(S \mid \nu, V) = c(\nu, V) \exp\left(\frac{\nu - d - 1}{2} \log(\det(S)) - \sum_{i,j} V_{i,j} S_{i,j}\right)$$

is characterized as the distribution of maximum entropy in the support $S > 0$, given the expected value of the elements and log-determinant of matrix S. That is, writing Γ' for the digamma function,

$$\mathrm{E}(S_{i,j}) = V_{i,j} \;\;,\;\; \mathrm{E}(\log(\det(S))) = \sum_{k=1}^{d} \Gamma'\left(\frac{\nu - k + 1}{2}\right)$$

The Dirichlet distribution

$$f(x \mid \theta) = c(\theta) \exp\left(\sum_{k=1}^{m} (\theta_k - 1) \log(x_k)\right)$$

is characterized as the distribution of maximum entropy in the simplex support, $x \geq 0 \mid \mathbf{1}'x = 1$, given the expected values of the log-coordinates, $\mathrm{E}(\log(x_k))$. In this parameterization, $\mathrm{E}(x_k) = \theta_k$.

Jeffrey's Rule:

Richard Jeffrey considered the problem of updating an old probability distribution, q, to a new distribution, p, given new constraints on the probabilities of a partition, that is,

$$\sum_{i \in S_k} p_i = \alpha_k \;,\;\; \sum_k \alpha_k = 1 \;,\;\; S_1 \cup \ldots \cup S_m = \{1, \ldots n\} \;,\;\; S_l \cap S_k = \emptyset,\; l \neq k$$

His solution to this problem, known as the *Jeffrey's rule*, coincides with the minimum information divergence distribution, relative to q, given the new constraints. This solution can be expressed analytically as

$$p_i = \alpha_k q_i / \sum_{j \in S_k} q_j, \;\; k \mid i \in S_k$$

4.5. Fisher's Metric and Jeffreys' Prior

In this section the Fisher Information Matrix is defined and used to obtain the geometrically invariant Jeffreys' prior distributions. These distributions also have interesting asymptotic properties concerning the representation of vague or no information. The properties of Fisher's metric discussed in this section are also needed to establish further asymptotic results in the next section.

The Fisher Information Matrix, $J(\theta)$, is defined as minus the expected Hessian of the log-likelihood. Under appropriate regularity conditions, the *information geometry* is defined by the metric in the parameter space given by the Fisher information matrix, that is, the geometric length of a curve is computed integrating the form $dl^2 = d\theta' J(\theta) d\theta$.

Lemma: The Fisher information matrix can also be written as the covariance matrix of the gradient of the same likelihood, *i.e.*,

$$J(\theta) \equiv -\,\mathrm{E}_{\mathcal{X}} \frac{\partial^2 \log p(x \mid \theta)}{\partial\, \theta^2} = \mathrm{E}_{\mathcal{X}} \left(\frac{\partial\, \log p(x \mid \theta)}{\partial\, \theta}\, \frac{\partial\, \log p(x \mid \theta)}{\partial\, \theta}\right)$$

Proof:

$$\int_{\mathcal{X}} p(x\,|\,\theta)dx = 1 \Rightarrow \int_{\mathcal{X}} \frac{\partial\, p(x\,|\,\theta)}{\partial\,\theta}dx = 0 \Rightarrow$$
$$\int_{\mathcal{X}} \frac{\partial\, p(x\,|\,\theta)}{\partial\,\theta}\frac{p(x\,|\,\theta)}{p(x\,|\,\theta)}dx = \frac{\partial\, \log p(x\,|\,\theta)}{\partial\,\theta}p(x\,|\,\theta)dx = 0$$

differentiating again relative to the parameter,

$$\int_{\mathcal{X}} \left(\frac{\partial^{\,2} \log p(x\,|\,\theta)}{\partial\,\theta^2}p(x\,|\,\theta) + \frac{\partial\, \log p(x\,|\,\theta)}{\partial\,\theta}\,\frac{\partial\, p(x\,|\,\theta)}{\partial\,\theta} \right) dx = 0$$

observing that the second term can be written as

$$\int_{\mathcal{X}} \frac{\partial\, \log p(x\,|\,\theta)}{\partial\,\theta}\,\frac{\partial\, p(x\,|\,\theta)}{\partial\,\theta}\,\frac{p(x\,|\,\theta)}{p(x\,|\,\theta)}dx = \int_{\mathcal{X}} \frac{\partial\, \log p(x\,|\,\theta)}{\partial\,\theta}\,\frac{\partial\, \log p(x\,|\,\theta)}{\partial\,\theta}\,p(x\,|\,\theta)dx$$

we obtain the lemma.

Harold Jeffreys used the Fisher metric to define a class of prior distributions, proportional to the determinant of the information matrix,

$$p(\theta) \propto |J\,(\theta)|^{1/2}$$

Lemma: Jeffreys' priors are geometric objects in the sense of being invariant by a continuous and differentiable change of coordinates in the parameter space, $\eta = f(\theta)$. The proof follows pp. 41–54 of [100]:

Proof:

$$J(\theta) = \left[\frac{\partial\,\eta}{\partial\,\theta}\right] J(\eta) \left[\frac{\partial\,\eta}{\partial\,\theta}\right]' \text{ , hence}$$
$$|J\,(\theta)|^{1/2} = \left|\frac{\partial\,\eta}{\partial\,\theta}\right| \, |J\,(\eta)|^{1/2} \text{ , and}$$
$$|J\,(\theta)|^{1/2}\; d\theta = |J\,(\eta)|^{1/2}\; d\eta \,. \text{ Q.E.D.}$$

Example: For the multinomial distribution,

$$p(y\,|\,\theta) = n! \textstyle\prod_{i=1}^{m} \theta_i^{x_i} \,\Big/\, \prod_{i=1}^{m} x_i! \;\;,\;\; \theta_m = 1 - \sum_{i=1}^{m-1} \theta_i \;,\; x_m = n - \sum_{i=1}^{m-1} x_i$$
$$L = \log p(\theta\,|\,x) = \textstyle\sum_{i=1}^{m} x_i \log \theta_i$$
$$\frac{\partial^{\,2} L}{(\partial\,\theta_i)^2} = -\frac{x_i}{\theta_i^2} + \frac{x_m}{\theta_m^2} \;,\;\; \frac{\partial^{\,2} L}{\partial\,\theta_i \partial\,\theta_j} = -\frac{x_m}{\theta_m^2} \;,\;\; i,j = 1 \ldots m-1$$
$$-\,\mathrm{E}_{\mathcal{X}} \frac{\partial^{\,2} L}{(\partial\,\theta_i)^2} = \frac{n}{\theta_i} + \frac{n}{\theta_m} \;,\;\; -\,\mathrm{E}_{\mathcal{X}} \frac{\partial^{\,2} L}{\partial\,\theta_i \partial\,\theta_j} = \frac{n}{\theta_m}$$
$$|J(\theta)| = (\theta_1 \theta_2 \ldots \theta_m)^{-1} \;\;,\;\; p(\theta) \propto (\theta_1 \theta_2 \ldots \theta_m)^{-1/2}$$
$$p(\theta\,|\,x) \propto \theta_1^{x_1 - 1/2} \theta_2^{x_2 - 1/2} \ldots \theta_m^{x_m - 1/2}$$

In general Jeffrey's priors are not minimally informative in any sense. However, in pp. 41–54 of [100], Zellner gives the following argument (attributed to Lindley) to present Jeffreys' priors as "knowing little" in the sense of being asymptotically minimally informative. The following equations give several

definitions related to the concept of information gain, that is expressed as the prior average information associated with an observation minus the prior information measure: $I(\theta)$—the information measure of $p(x \mid \theta)$, A—the prior average information associated with an observation, G—the information gain, and G_a—the asymptotic information gain.

$$I(\theta) = \int p(x \mid \theta) \log p(x \mid \theta) dx \; ; \quad A = \int I(\theta) p(\theta) d\theta$$

$$G = A - \int p(\theta) \log p(\theta) d\theta \; ; \quad G_a = \int p(\theta) \sqrt{n \, |J(\theta)|} d\theta - \int p(\theta) \log p(\theta) d\theta$$

Although Jeffreys' priors in general do not maximize the information gain, G, the asymptotic convergence results presented in the next section imply that Jeffrey's priors maximize the asymptotic information gain, G_a. For further details and generalizations, see [101–110].

Comparing the several versions of noninformative priors in the multinomial example, one can say that Jeffreys' prior "discounts" half an observation of each kind, while the maxent prior discounts one full observation, and the flat prior discounts none. Similarly, slightly different versions of uninformative priors for the multivariate normal distribution are shown in [106]. This situation leads to the possible criticism stated by Berger in p. 89 of [104]:

> *Perhaps the most embarrassing feature of noninformative priors, however, is simply that there are often so many of them.*

One response to this criticism, to which Berger explicitly subscribes in p. 90 of [104], is that

> *It is rare for the choice of a noninformative prior to markedly affect the answer... so that any reasonable noninformative prior can be used. Indeed, if the choice of noninformative prior does have a pronounced effect on the answer, then one is probably in a situation where it is crucial to involve subjective prior information.*

The robustness of the inference procedures to variations on the form of the uninformative prior can be tested using sensitivity analysis, as discussed in Section 4.7 of [104]. For alternative approaches on robustness and sensitivity analysis based on paraconsistent logic, see [4,5].

4.6. Posterior Asymptotic Convergence

Posterior convergence constitutes the principal mechanism enabling information acquisition or learning in Bayesian statistics. Arguments based on relative information, $I(p, q)$, can be used to prove fundamental results concerning posterior distribution asymptotic convergence. This section presents two of these fundamental results, following Appendix B of [96].

Theorem: Posterior Consistency for Discrete Parameters:

Consider a model where $f(\theta)$ is the prior in a discrete parameter space, $\Theta = \{\theta^1, \theta^2, \ldots\}$, $X = [x^1, \ldots x^n]$ is a series of observations, and the posterior is given by

$$f(\theta^k \mid X) \propto f(\theta^k) \, p(X \mid \theta^k) = f(\theta^k) \, \textstyle\prod_{i=1}^{n} p(x^i \mid \theta^k)$$

Further, assume that this model has a unique vector parameter, θ^0, giving the best approximation for the "true" predictive distribution $g(x)$, in the sense that it minimizes the relative information

$$\begin{aligned} \{\theta^0\} &= \arg\min_k I\left(g(x), p(x \mid \theta^k)\right) \\ I\left(g(x), p(x \mid \theta^k)\right) &= \int_{\mathcal{X}} g(x) \log\left(\frac{g(x)}{p(x \mid \theta^k)}\right) dx = \mathrm{E}_{\mathcal{X}} \log\left(\frac{g(x)}{p(x \mid \theta^k)}\right) \end{aligned}$$

Then,

$$\lim_{n\to\infty} f(\theta^k \mid X) = \delta(\theta^k, \theta^0)$$

Heuristic Argument: Consider the logarithmic coefficient

$$\log\left(\frac{f(\theta^k \mid X)}{f(\theta^0 \mid X)}\right) = \log\left(\frac{f(\theta^k)}{f(\theta^0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(x^i \mid \theta^k)}{p(x^i \mid \theta^0)}\right)$$

The first term is a constant, and the second term is a sum which terms have all negative expected (relative to x, for $k \neq 0$) value since, by our hypotheses, θ^0 is the unique argument that minimizes $I(g(x), p(x \mid \theta^k))$. Hence, (for $k \neq 0$), the right hand side goes to minus infinite as n increases. Therefore, at the left hand side, $f(\theta^k \mid X)$ must go to zero. Since the total probability adds to one, $f(\theta^0 \mid X)$ must go to one, QED.

We can extend this result to continuous parameter spaces, assuming several regularity conditions, like continuity, differentiability, and having the argument θ^0 as an interior point of Θ with the appropriate topology. In such a context, we can state that, given a pre-established small neighborhood around θ^0, like $C(\theta^0, \epsilon)$ the cube of side size ϵ centered at θ^0, this neighborhood concentrates almost all mass of $f(\theta \mid X)$, as the number of observations grows to infinite. Under the same regularity conditions, we also have that Maximum a Posteriori (MAP) estimator is a consistent estimator, *i.e.*, $\widehat{\theta} \to \theta^0$.

The next results show the convergence in distribution of the posterior to a Normal distribution. For that, we need the Fisher information matrix identity from the last section.

Theorem: Posterior Normal Approximation:

The posterior distribution converges to a Normal distribution with mean θ^0 and precision $nJ(\theta^0)$.

Proof (heuristic)**:** We only have to write the second order log-posterior Taylor expansion centered at $\widehat{\theta}$,

$$\begin{aligned} \log f(\theta \mid X) &= \log f(\widehat{\theta} \mid X) + \frac{\partial \log f(\widehat{\theta} \mid X)}{\partial \theta}(\theta - \widehat{\theta}) \\ &+ \frac{1}{2}(\theta - \widehat{\theta})' \frac{\partial^2 \log f(\widehat{\theta} \mid X)}{\partial \theta^2}(\theta - \widehat{\theta}) + \mathcal{O}(\theta - \widehat{\theta})^3 \end{aligned}$$

The term of order zero is a constant. The linear term is null, for $\widehat{\theta}$ is the MAP estimator at an interior point of Θ. The Hessian in the quadratic term is

$$H(\widehat{\theta}) = \frac{\partial^2 \log f(\widehat{\theta} \mid X)}{\partial \theta^2} = \frac{\partial^2 \log f(\widehat{\theta})}{\partial \theta^2} + \sum_{i=1}^{n} \frac{\partial^2 \log p(x^i \mid \widehat{\theta})}{\partial \theta^2}$$

The Hessian is negative definite, by the regularity conditions, and because $\widehat{\theta}$ is the MAP estimator. The first term is constant, and the second is the sum of n i.i.d. random variables. At the other hand we have

already shown that the MAP estimator, and also that all the posterior mass concentrates around θ^0. We also see that the Hessian grows (in average) linearly with n, and that the higher order terms can not grow super-linearly. Also for a given n and $\theta \to \widehat{\theta}$, the quadratic term dominates all higher order terms. Hence, the quadratic approximation of the log-posterior in increasingly more precise, Q.E.D.

5. Final Remarks

The objections raised by Spencer-Brown against probability and statistics, analyzed in Sections 1 and 2, are somewhat simplistic and stereotypical, possibly explaining why they had little influence outside a close circle of admirers, most of them related to the radical constructivism movement. However, arguments very similar to those used to demystify Spencer-Brown's misconceptions and elucidate its misunderstandings, reappear in more subtle or abstract forms in the analysis of far more technical matters like, for example, the use and interpretation of prior and posterior distributions in Bayesian statistics.

In this article, entropy is presented as a cornerstone concept for the precise analysis and a key idea for the correct understanding of several important topics in probability and statistics. This understanding should help to clear the way for establishing Bayesian statistics as a preferred tool for scientific inference in mainstream cognitive constructivism.

Acknowledgements

The author is grateful for the support of the Department of Applied Mathematics of the Institute of Mathematics and Statistics of the University of São Paulo, FAPESP—Fundação de Amparo à Pesquisa do Estado de São Paulo, and CNPq—Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant PQ-306318-2008-3). The author is also grateful for the helpful discussions with several of his professional colleagues, including Carlos Alberto de Bragança Pereira, Fernando Bonassi, Luis Esteves, Marcelo de Souza Lauretto, Rafael Bassi Stern, Sergio Wechsler and Wagner Borges.

References

1. Borges, W.; Stern, J.M. The rules of logic composition for the bayesian epistemic e-values. *Log. J. IGPL* **2007**, *15*, 401–420.
2. Pereira, C.A.B.; Stern, J.M. Evidence and credibility: Full bayesian significance test for precise hypotheses. *Entropy* **1999**, *1*, 69–80.
3. Pereira, C.A.B.; Wechsler, S.; Stern, J.M. Can a significance test be genuinely bayesian? *Bayesian Anal.* **2008**, *3*, 79–100.
4. Stern, J.M. Significance Tests, Belief Calculi, and Burden of Proof in Legal and Scientific Discourse. *Laptec-2003, Frontiers in Artificial Intelligence and Its Applications*; ISO Press: Amsterdam, The Netherlands, 2003; Volume 101, pp. 139–147.
5. Stern, J.M. Paraconsistent Sensitivity Analysis for Bayesian Significance. Tests SBIA'04, In *Lecture Notes Artificial Intelligence*; Goebel, R., Siekmann, J., Wahlster, W., Eds.; Springer: Heidelberg , Germany, 2004; Volume 3171, pp. 134–143.

6. Stern, J.M. *Language, Metaphor and Metaphysics*: *The Subjective Side of Science*; Technical Report MAC-IME-USP-06-09; Department of Statistical Science, University College: London, UK, 2006.
7. Stern, J.M. Cognitive constructivism, eigen-solutions, and sharp statistical hypotheses. *Cybern. Hum. Knowing* **2007**, *14*, 9–36.
8. Stern, J.M. Language and the self-reference paradox. *Cybern. Hum. Knowing* **2007**, 14, 71–92.
9. Stern, J.M. *Complex Structures*, *Modularity and Stochastic Evolution*; Technical Report IME-USP-MAP-07-01; University of Sao Paulo, Sao Paulo, Brazil, 2007.
10. Stern, J.M. Decoupling, sparsity, randomization, and objective bayesian inference. *Cybern. Hum. Knowing* **2008**, *15*, 49–68.
11. Stern, J.M. Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses. Presented at MaxEnt 2008, The 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Boraceia, Sao Paulo, Brazil, 2008.
12. Stern, J.M. The Living and Intelligent Universe. In *Proceeding of MBR09-The Internaternational Conference on Model-Based Reasoning in Science and Technology*, Unicamp, Brazil, 2009.
13. von Foerster, H. *Understanding Understanding*: *Essays on Cybernetics and Cognition*; Springer Verlag: New York, NY, USA, 2003.
14. Bernardo, G.G.; Lauretto, M.S.; Stern, J.M. The Full Bayesian Significance Test form Symmetry in Contingency Tables. In *Proceeding of 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Chamonix, France, July 4–9, 2010,
15. Chakrabarty, D. CHASSIS-Inverse Modelling of Relaxed Dynamical Systems. In *Proceedings of the 18th World IMACS MODSIM Congress*, Cairns, Australia, 13–17 July 2009.
16. Johnson, R.; Chakrabarty, D.; O'Sullivan, E.; Raychaudhury, S. Comparing X-ray and dynamical mass profiles in the early-type galaxy NGC 4636. *Astrophys. J.* **2009**, *706*, 980–994.
17. Loschi, R.H.; Monteiro, J.V.D.; Rocha, G.H.M.A.; Mayrink, V.D. Testing and estimating the non-disjunction fraction in meiosis I using reference priors. *Biom. J.* **2007**, *49*, 824–839.
18. Madruga, M.R; Esteves, L.G.; Wechsler, S. On the bayesianity of pereira-stern tests. *Test* **2001**, *10*, 291–299.
19. Rifo, L.L.; Torres, S. Full bayesian analysis for a class of jump-diffusion models. *Comm. Stat. Theor. Meth.* **2009**, *38*, 1262–1271.
20. Rodrigues, J. Full bayesian significance test for zero-inflated distributions. *Comm. Stat. Theor. Meth.* **2006**, *35*, 299–307.
21. Colla, E.; Stern, J.M. Sparse factorization methods for inference in bayesian networks. *AIP Conf. Proc.* **2008**, *1073*, 136–143.
22. Hacking, I. Telepathy: Origins of Randomization in Experimental Design. *Isis* **1988**, *79*, 427–451.
23. Peirce, C.S.; Jastrow, J. On small differences of sensation. *Memoirs of the National Academy of Sciences*; National Academies Press: Washingtion, DC, USA, **1884**, *3*, 75–83.
24. Spencer-Brown, G. Statistical significance in psychical research. *Nature* **1953**, *172*, 154–156.
25. Spencer-Brown, G. Answer to soal *et al*. *Nature* **1953**, *172*, 594–595.
26. Spencer-Brown, G. *Probability and Scientific Inference*; Longmans Green: London, UK, 1957.
27. Spencer-Brown, G. *Laws of Form*; Allen and Unwin: London, UK, 1969.

28. Carnielli, W. Formal Polynomials and the Laws of Form. In *Dimensions of Logical Concepts*; Béziau, J.Y., Costa-Leite, A., Eds.; UNICAMP: Campinas, Brazil, 2009.
29. Edwards, A.W.F. *Cogwheels of the Mind: The Story of Venn Diagrams*; The Johns Hopkins University Press: Baltimore, MD, USA, 2004.
30. Kauffman, L.H. The mathematics of charles sanders peirce. *Cybern. Hum. Knowing* **2001**, *8*, 79–110.
31. Kauffman, L.H. Laws of Form: An Exploration in Mathematics and Foundations, 2006. Available at: http:// www.math.uic.edu/kauffman/Laws.pdf (accessed on 1 April 2011)
32. Meguire, P. Discovering boundary algebra: A simple notation for boolean algebra and the truth functions. *Int. J. Gen. Sys.* **2003**, *32*, 25–87.
33. Peirce, C.S. A Boolean Algebra with One Constant. In *Collected Papers of Charles Sanders Peirce*; Hartshorne, C., Weiss, P., Burks, A., Eds.; InteLex: Charlottesville, VA, USA, 1992.
34. Sheffer, H.M. A Set of five independent postulates for boolean algebras, with application to logical constants. *Trans. Amer. Math. Soc.* **1913**, *14*, 481–488.
35. Flew, A. Probability and statistical inference by G.Spencer-Brown (review). *Phil. Q.* **1959**, *9*, 380–381.
36. Falk, R.; Konold, C. Making sense of randomness: Implicit encoding as a basis for judgment. *Psychol. Rev.* **1997**, *104*, 301–318.
37. Falk, R.; Konold, C. Subjective randomness. In *Encyclopedia of Statistical Sciences*, 2nd ed.; Wiley: New York, NY, USA, 2005; Volume 13, pp. 8397–8403.
38. Good, I.J. Probability and statistical inference by G.Spencer-Brown (review). *Br. J. Philos. Sci.* **1958**, *9*, 251–255.
39. Mundle, C.W.K. Probability and statistical inference by G.Spencer-Brown (review). *Philosophy* **1959**, *34*, 150–154.
40. Atkins, P.W. *The Second Law*; The Scientific American Books: New York, NY, USA, 1984.
41. Attneave, E. *Applications of Information Theory to Psychology*: *A Summary of Basic Concepts, Methods*, *and Results*; Holt, Rinehart and Winston: New York, NY, USA, 1959.
42. Dugdale, J.S. *Entropy and Its Physical Meaning*; Taylor and Francis: London, UK, 1996.
43. Krippendorff, K. *Information Theory*: *Structural Models for Qualitative Data* (*Quantitative Applications in the Social Sciences V.62.*); Sage: Beverly Hills, CA, USA, 1986.
44. Tarasov, L. *The World Is Built on Probability*; MIR: Moscow, Russia, 1988.
45. Kapur, J.N. *Maximum Entropy Models in Science and Engineering*; John Wiley: New Delhi, India, 1989.
46. Rissanen, J. *Stochastic Complexity in Statistical Inquiry*; World Scientific: New York, NY, USA, 1989.
47. Wallace, C.S. *Statistical and Inductive Inference by Minimum Message Length*; Springer: NewYork, NY, USA, 2005.
48. Tribble, C.G. Industry-sponsored negative trials and the potential pitfalls of post hoc analysis. *Arch. Surg.* **2008**, *143*, 933–934.
49. Wang, R.; Lagakos, S.W.; Ware, J.H.; Hunter, D.J.; Drazen, J.M. Statistics in medicine-reporting of subgroup analyses in clinical trials. *New Engl. J. Med.* **2007**, *357*, 2189–2194.

50. Scott, C.G. Spencer-brown and probability: A critique. *J. Soc. Psych. Res.* **1958**, *39*, 217–234.
51. Soal, S.G.; Stratton, F.J.; Thouless, R.H. Statistical significance in psychical research. *Nature* **1958**, *172*, 594.
52. Atmanspacher, H. Non-physicalist physical approaches. guest editorial. *Mind Matter* **2005**, *3*, 3–6.
53. Ehm, W. Meta-analysis of mind-matter experiments: A statistical modeling perspective. *Mind Matter* **2005**, *3*, 85–132.
54. Henning, C. *Falsification of Propensity Models by Statistical Tests and the Goodness-of-Fit Paradox*. Technical Report no. 304; Department of Statistical Science, University College, London. 2006.
55. Kaptchuk, T.J., Kerr, C.E. Commentary: Unbiased divination, unbiased evidence, and the patulin clinical trial. *Int. J. Epidemiol.* **2004**, *33*, 247–251.
56. Utts, J. Replication and meta-analysis in parapsychology. *Stat. Sci.* **1991**, *6*, 363–403.
57. Wassermann, G.D. Some comments on the methods and statements in parapsychology and other sciences. *Br. J. Philos. Sci.* **1955**, *6*, 122–140.
58. Bonassi, F.V.; Stern, R.B.; Wechsler, S. The gambler's fallacy: a bayesian approach. *AIP Conf. Proc.* **2008**, *1073*, 8–15.
59. Bonassi, F.V.; Nishimura, R.; Stern, R.B. In defense of randomization: A subjectivist bayesian approach. *AIP Conf. Proc.* **2009**, *1193*, 32–39.
60. Dehue, T. Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *Isis* **1997**, *88*, 653–673.
61. Hammersley, J.M.; Handscomb, D.C. *Monte Carlo Methods*; Chapman and Hall: London, UK, 1964.
62. Ripley, B.D. *Stochastic Simulation*; Wiley: New York, NY, USA, 1987.
63. Marsaglia, G. Random numbers fall mainly in the planes. *Proc. Natl. Acad. Sci.* **1968**, *61*, 25–28.
64. Boyar, J. Inferring sequences produced by pseudo-random number generators. *J. ACM* **1989**, *36*, 129–141.
65. Matsumoto, M.; Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Model. Comput. Simul.* **1998**, *8*, 3–30.
66. Matsumoto, M.; Kurita, Y. Twisted GFSR generators. *ACM Trans. Model. Comput. Simul.* **1992**, *2*, 179–194.
67. Abelson, R.P. *Statistics as Principled Argument*; LEA: Hillsdale, NJ, USA, 1995.
68. Matoušek, J. *Geometric Discrepancy*; Springer: Berlin, Germany, 1991.
69. Günther, M.; Jüngel, A. *Finanzderivate mit MATLAB. Mathematische Modellierung und Numerische Simulation*; Vieweg Verlag: Wiesbaden, Germany, 2003; p. 117.
70. Merkel, R. Analysis and Enhancements of Adaptive Random Testing. Ph.D. Thesis, Swinburne University of Technology in Melbourne, Melbourne, Australia, 2005.
71. Ökten, G. Contributions to the Theory of Monte Carlo and Quasi monte Carlo Methods. Ph.D. Thesis, Clearmont University, Clearmont, CA, USA, 1999.
72. Sen, S.K.; Samanta, T.; Reese, A. Quasi versus pseudo random generators: Discrepancy, complexity and integration-error based comparisson. *Int. J. Innov. Comput. Inform. Control* **2006**, *2*, 621–651.

73. Morokoff, W.J. Generating quasi-random paths for stochastic processes. *SIAM Rev.* **1998**, *40*, 765–788.
74. Zabell, S.L. The Quest for Randomness and its Statistical Applications. In *Statistics for the Twenty-First Century*; Gordon, E., Gordon, S., Eds.; Mathematical Association of America: Washington, DC, USA, 1992.
75. Gell'Mann, M. *The Quark and the Jaguar*: *Adventures in the Simple and the Complex*; W. H. Freeman: New York, NY, USA, 1994.
76. Lopes, L.L. Doing the Impossible: a note on induction and the experience of randomness. *J. Exp. Psychol. Learn. Mem. Cognit.* **1982**, *8*, 626–636.
77. Lopes, L.L.; Oden, G.C. Distinguishing between random and nonrandom events. *J. Exp. Psychol. Learn. Mem. Cognit.* **1987**, *13*, 392–400.
78. Tversky, Y; Kahneman, D. Belief in the law of small numbers. *Psychol. Bull.* **1971**, *76*, 105–110.
79. Piaget, J.; Inhelder, B. *The Origin of the Idea of Chance in Children*; Leake, L., Burrell, E., Fishbein, H.D., Eds.; Norton: New York, NY, USA, 1975.
80. Chaitin, G.J. Randomness and mathematical proof. *Sci. Amer.* **1975**, *232*, 47–52.
81. Chaitin, G.J. Randomness in arithmetic. *Sci. Amer.* **1988**, *259*, 80–85.
82. Kac, M. What is random? *Amer. Sci.* **1983**, *71*, 405–406.
83. Kolmogorov, A.N. Three approaches to the quantitative definition of information. *Probl. Inform. Transm.* **1965**, *1*, 1–7.
84. Martin-Löf, E. The definition of random sequences. *Inform. Contr.* **1966**, *9*, 602–619.
85. Martin-Löf, P. Algorithms and randomness. *Int. Statist. Inst.* **1969**, *37*, 265–272.
86. Csiszar, I. Information Measures. In *Proceedings of the 7th Prage Conferences of Information Theory*, Prague, Czech Republic, 1974; Volume 2, pp. 73–86.
87. Khinchin, A.I. *Mathematical Foundations of Information Theory*; Dover: New York, NY, USA, 1953.
88. Renyi, A. On Measures of Entropy and Information. In Proceedings of the 4th Berkeley Symposium on Mathematical and Statistical Problems. Statistical Laboratory of the University of California, Berkeley, June 20–July 30, 1960; University of California Press: Berkeley, CA, USA, 1961; Volume VI, pp. 547–561.
89. Renyi, A. *Probability Theory*; North-Holland: Amsterdam, the Netherlands, 1970.
90. Gokhale, D.V. Maximum Entropy Characterization of Some Distributions. In *Statistical Distributions in Scientific Work*; Patil, G.P., Kotz, G.P., Ord, J.K., Eds.; Springer: Berlin, Germany, 1975; Volume 3, pp. 299–304.
91. Censor, Y.; Zenios, S. *Introduction to Methods of Parallel Optimization*; IMPA: Rio de Janeiro, Brazil, 1994.
92. Censor, Y.; Zenios, S.A. *Parallel Optimization: Theory, Algorithms, and Applications*; Oxford University Press: New York, NY, USA, 1997.
93. Elfving, T. On some methods for entropy maximization and matrix scaling. *Linear Algebra Appl.* **1980**, *34*, 321–339.
94. Fang, S.C.; Rajasekera, J.R.; Tsao, H.S.J. *Entropy Optimization and Mathematical Programming*; Kluwer: Dordrecht, The Netherlands, 1997.

95. Iusem, A.N.; Pierro, A.R. De Convergence results for an accelerated nonlinear cimmino algorithm. *Numer. Math.* **1986**, *46*, 367–378.
96. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd ed; Chapman and Hall/CRC: New York, NY, USA, 2003.
97. Caticha, A. Lectures on Probability, Entropy and Statistical Physics. Presented at MaxEnt 2008, The 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Boracéia, São Paulo, Brazil, 2008.
98. Tribus, M.; McIrvine, E.C. Energy and information. *Sci. Amer.* **1971**, *224*, 178–184.
99. Garcia, M.V.P.; Humes, C.; Stern, J.M. Generalized line criterion for gauss seidel method. *J. Comput. Appl. Math.* **2002**, *22*, 91–97.
100. Zellner, A. *Introduction to Bayesian Inference in Econometrics*; Wiley: New York, NY, USA, 1971.
101. Amari, S.I.; Barndorff-Nielsen, O.E.; Kass, R.E.; Lauritzen, S.L.; Rao, C.R. Differential Geometry in Statistical Inference. *IMS Lecture Notes Monograph*; Institute of Mathematical Statistics: Hayward, CA, USA, 1987; Volume 10.
102. Amari, S.I. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2007.
103. Berger, J.O.; Bernardo, J.M. On the Development of Reference Priors. In *Bayesian Statistics 4*; Bernardo, J.M., Berger, J.O., Lindley, D.V., Smith, A.F.M., Eds.; Oxford University Press: Oxford, UK, 1992, 35–60.
104. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed; Springer: New York, NY, USA, 1993.
105. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; Wiley: New York, NY, USA, 2000.
106. DeGroot, M.H. *Optimal Statistical Decisions*; McGraw-Hill: New York, NY, USA, 1970.
107. Hartigan, J.A. *Bayes Theory*; Springer: New York, NY, USA, 1983.
108. Jeffreys, H. *Theory of Probability*, 3rd ed.; Clarendon Press: Oxford, UK, 1961.
109. Scholl, H. Shannon optimal priors on independent identically distributed statistical experiments converge weakly to Jeffreys' prior. *Test* **1998**, *7*, 75–94.
110. Zhu, H. *Information Geometry, Bayesian Inference, Ideal Estimates and Error Decomposition*; Santa Fe Institute: Santa Fe, NM, USA, 1998.