Info-Metrics in the Natural Sciences and Its Implications for the Social Sciences

Amos Golan (American University and Info-Metrics)

The study of information, the philosophy of information and entropy as a method of inference provides a unified approach to inference and learning across the disciplines.

The Info-Metrics Institute promotes the study of information, information processing and optimal decision rules based on efficient use of information.

1. Workshop objectives: (general)

   In this (inherently interdisciplinary) workshop we wish to:
   
   A. Lay out the basics of info-metrics in the natural sciences.
   
   B. Study the state of information-theoretic estimation and data analysis in some areas of the natural sciences.
   
   C. Discuss some of the basic philosophy of information and information processing.
   
   D. Discuss the implications of info-metrics in the natural sciences to the social sciences in general (and to economics, econometrics and finance).

   Among the basic questions we want to reflect on are:

   A. What is information? Is all information the same?
   
   B. What is the basic unit of analysis? What are the quantities we observe? What is it we are trying to understand?
   
   C. How to do inference with limited information?
   
   D. How can we understand a complex system?
   
   E. Is a unified approach to inference and learning necessary/useful, and what other unification approaches compete with IT?

   We want to study how (some of) these issues are treated in the natural sciences, and to think what we can learn from this in terms of new directions and implementations toward information processing in the social sciences.

2. A Brief Background

   Our objective in science is to understand and explain the world around us (or a specific system/phenomenon of interest) from the available information. This information (data) is a mix of evidence, hard data, theories, intuition, conjectures, priors and other “soft” data.) So, can we find the absolute truth (in our life time)?

   All learning, information gathering and information processing, is based on limited knowledge, both a-priori and data, from which a larger “truth” must be inferred. To learn about the true state of the world that generated the observed data, we use statistical models. Since we will never know the true state of the world, we generally focus, in statistical sciences, on recovering information about the complete probability distribution, which represents the ultimate truth in our model. Therefore, all estimation and inference problems are translations of limited information about the probability density function (pdf) toward a greater knowledge of that pdf.

   But a main problem we face: Information and data are imperfect.

   Therefore, extracting the correct information from these data, without imposing structure or theory that often cannot be confirmed, is a tough task. (We often see numerous interpretations resulting from the analysis of a single data set.) Why?

   Simply stated: Unless we impose specific structure due to assumptions we cannot confirm, the estimation problem becomes inherently underdetermined: There are more unknown quantities than observed (known) quantities, leading to infinitely many possible solutions (interpretations or theories) that are consistent with the same data.

   Which solution to choose? How to go about it? How can we guarantee that only the relevant information is extracted? How can we assess that information? How can we process this information?

   Entropy, Generalized Entropy, Information Theory and Maximum Entropy are the answers.

   But How?

   These are the basic questions we study today.
The concept of entropy had its origins in the 19th century in the discovery of thermodynamics (Carnot, Clausius, Kelvin) and statistical mechanics (Maxwell, Boltzmann, Gibbs). A series of developments starting around the middle of the 20th century (mostly due to Shannon and Jaynes) liberated the concept of entropy from its origins in physics and elevated it into a general purpose tool for processing information. Thus was born the old Method of Maximum Entropy – MaxEnt.

In a parallel line of research, particularly throughout the latter half of the 20th century, Bayesian inference enjoyed a remarkably sustained period of expansion. The two methods of inference flourished side by side and to a large extent quite independently of each other; indeed, entropic methods do not yet enjoy widespread acceptance within the orthodox Bayesian community. The connection between them has been an endless source of controversy and even their compatibility has been repeatedly brought into question.

Further developments in entropic inference extending well into the early 21st century have, however, culminated in the complete unification of entropic and Bayesian inference methods. Indeed, the new Method of Maximum Entropy includes the old MaxEnt, all Bayesian methods, and the general subject of large deviations as special cases.

The consequences of these developments are potentially enormous both for statistics (entropic priors, model selection, experimental design, etc.) and for science in general. In physics, for example, it is well known that the laws of thermodynamics can be derived from entropic methods. What might at first sight be unexpected is that entropic principles can also be used to derive other fundamental laws such as Newton’s $F = ma$ in classical mechanics, and the Schrödinger equation in quantum mechanics. Such an unusual perspective yields new insights about several central concepts in physics such as the nature of time and the phase of the wave function, but perhaps more importantly it suggests that the laws of physics might not be as fundamental as normally believed. Rather than being laws of nature they might be highly effective rules to process information about the world – just like any other science.

The examples I discuss are drawn from physics but this is a reflection of my own personal tastes. This research program is driven by the conviction that the laws of inference are of universal applicability – once the appropriate relevant information is codified into a statistical model there is no reason why the methods of entropic inference should not find application in other fields.
Maximal Entropy Inference of Oncogenicity
Raphael D. Levine (The Hebrew U)

Systems Biology is of much interest in its own right. It is also ideally placed as a bridge linking the applications of information theoretic methods between the natural sciences and the social sciences. This is for at least two reasons. One to be discussed more below is the question of ‘whose entropy’. The other consideration is that the ‘invariance’ so commonly taken for granted in Physics, is an invariance that can no longer be assumed in Biology and surely is questionable in the social sciences. This is the invariance that is simply expressed as all atoms of hydrogen are equivalent. For example not all T cells are obviously equivalent because T cells can learn from their previous history.

In physics and chemistry the physical entropy as introduced in thermodynamics is the same as the information theoretic, IT, entropy if and only if we are dealing with the distribution of quantum states. Computing the IT entropy using any other distribution, say the distribution of energy states, will lead to two not equal entropies. (By IT entropy I mean the expression that Shannon introduced on axiomatic grounds). Another manifestation of this problem is that a measured distribution need not be one of maximal IT entropy. A celebrated example is the Otto Stern experimental verification of the Maxwell-Boltzmann velocity distribution in a gas. The agreement was not acceptable to the high standards of Stern. Einstein then pointed out that what Stern measured is the flux distribution whereas a distribution of maximal IT entropy is the velocity distribution. Of course, using a suitable prior will fix the problem BUT the prior must come from somewhere. In physics and chemistry the prior comes from the axiom that the distribution of maximal IT entropy is the distribution of quantum states. This works rather well when the prior can be computed [1]. But it is a much too fine grained a description when we deal with biological molecules.

One option is to try to determine the prior from the experimental data, an approach pioneered in [2]. Another option is to try to reason out what would be a reasonable prior [3] and then seek to validate the choice. Key points about [3] and in general can be found in the discussion comments of Amos Golan.


Discussion of Raphael D. Levine’s “Maximal Entropy Inference of Oncogenicity”
Discussant: Amos Golan

The discussion will be at two (interrelated) levels:
I. The fundamentals (what is it that we are doing? What is (are) the basic quantity (quantities) that we maximize over? What do we observe? What information we have? …)
II. The more mathematical problem and some differences between the natural and the social sciences.

The Basics of the Model

A “Translated” Summary:

The Model
1. Pure model (Classical – Jaynes – Maximum Entropy)
2. Assume Normality \( \rightarrow \) sufficient to introduce two moments in the constraints
3. Using the Maximum Entropy principle \( \rightarrow \) Able to identify the (relative) contribution of each constraint (via the Lagrange multipliers)
4. Data are analyzed in terms of deviations from the means (no intercept)
5. The Lagrange multipliers reflect the information in the constraints \( \rightarrow \) represent the “normalized” covariance \( \rightarrow \) the correlation between the input (X) and the output (Y)
6. Original problem is over-determined/over-identified

The Empirical Results
1. A single constraint is most informative (in terms of entropy reduction) – almost 60%
2. A very small number of constraints are sufficient for good (“perfect”) prediction.

Contribution
1. Method: Converting an over-identified pure problem to a well posed problem; formulation as an IT-ME problem; connecting the Lagrange multipliers (estimated via Maximum Entropy) to the eigenvalues of the data and to the elements of the covariance (and correlation) and providing the informational interpretation.
2. Inference of oncogenicity from phosphorylation signaling (finding the main contributors to prediction) – a major new finding.

Main Discussion Points (Natural vs. Social Sciences):
1. The classical ME, ML and concentrated ME (Information and Lagrange multipliers)
2. Pure vs. noise (all data in the social sciences have noise)
3. Normality assumption for small samples
4. Sample vs. population

The Inference of Oncogenicity problem can be thought of (abstractly) as a “matrix balancing” problem. Examples of this class of problems include, variations of Levine’s mass-spectrometry problem, first order Markov transition probabilities, tomography, balancing social accounting (input-output) matrices, analysis of political survey, unordered multinomial choice problem, optimal management problems, image reconstruction, spectral analysis and more.

The basic problem was formulated as a pure ME problem. It is possible to generalize it (IT-GME). Basic advantages of the more general ME model is that it accommodates for the noise without added complexity (same number of Lagrange parameters/multipliers), outperforms the ME for small, complex and noisy data, computationally and statistically efficient and consistent with much of real world observed data (in social – and natural – sciences).

The Fundamentals: Discussion of the fundamentals – Basic points (see also Levine’s discussion):
- What is it that we are doing?
- What is (are) the basic quantity (quantities) that we maximize over?
- What do we observe?
- What information we have?
- What is the relationship between the observable and the basic quantity of interest?
- Micro vs. Macro level description
- What is the researcher’s objective? (Understanding the individuals’ interaction/behavior/exchange? Understanding the system as a whole? Causality…)
- What can we learn from the natural sciences and statistical physics?
- What are the (main) differences?
References:
2. Golan Discussion (a pdf version of the talk)
ApEn, approximate entropy, is a family of parameters and statistics quantifying regularity (orderliness) in serial data. ApEn has been developed both within a theoretical mathematical setting, especially to the classification of binary (finite and infinite) sequences, and also has seen numerous applications to multiple biological contexts. We consider the motivation for the initial ApEn development, from the study of inappropriate application of dynamical systems (“complexity”) algorithms to general time-series settings. Theoretically, ApEn is shown to extend both the rate of entropy for discrete state Markov Chains and the Kolmogorov-Sinai entropy for deterministic dynamical systems to settings in which those measures have no natural definition, e.g., the myriad composite models that likely underlie networks within systems biology. ApEn is scale invariant and model-independent, evaluates both dominant and subordinate patterns in data, and discriminates series for which clear feature recognition is difficult. ApEn is applicable to systems with at least 50 data points and to broad classes of models: it can be applied to discriminate both general classes of correlated stochastic processes, as well as noisy deterministic systems. Moreover, ApEn is complementary to spectral and autocorrelation analyses, providing effective discriminatory capability in instances in which the aforementioned measures exhibit minimal distinctions. Representative ApEn applications to diverse biomedical studies, including endocrinologic (hormonal secretory), heart rate, EEG, and neuronal (and Ca^{2+}) firing pattern time-series, are prominently featured. These include topics of broad public concern, including heart rate analyses of ‘near-SIDS’ infants, prediction of atrial and ventricular fibrillation, and gender- and age-related changes in cardiovascular dynamics in older subjects; epileptic seizure prediction; balance and stability changes as a marker of post-concussion recovery; and endocrine tumor detection. Furthermore, ApEn has been effectively applied to discriminate previously challenging data sets in more ‘bench’ or basic science studies, including e.g., respiration studies in neonatal piglets, differentiating coding from non-coding DNA regions, evaluating changes in connectivity between brain hemispheres, and discerning subtle yet highly significant changes to ion channel dynamics in single cell studies.
Comments on ApEnt by Steve Pincus
Essie Maasoumi: Emory University
April 27, 2010

- Given these relations to Cross Entropies, which transformations of ApEnt may provide more accessible metrics for assessing and testing CONDITIONAL hypotheses?
- Are APEnt methods "Invariant" to nonlinear data transformations in nonlinear settings?
- How can one accommodate hypotheses like symmetry, TR, and others with ApEnt?
- Can one accommodate "model based" conditioning methods which are very popular and meaningful in economics and Finance (see my slides)?
- VERY thought provoking indeed!
Comments on “Info-metrics on Medicine and Biology” by Steve Pincus

Ximing Wu*

In this presentation, Dr. Pincus introduced the concept of Approximate Entropy (ApEn) and provided many fascinating examples on how ApEn can be applied to various areas in Medicine and Biology. Some applications to economic times series are also discussed.

The efficacy of ApEn as a tool for discriminating analysis is clearly demonstrated through many examples. From the view of a potential user of this powerful tool, I would like to comment on some practical issues on ApEn. ApEn is a measure of irregularity that is indexed by two parameters: width of filter and blocklength of window. It is suggested that the former be set at 20% of standard deviation and the later be one or two. Alternatively, one might consider a rank-based ApEn. The advantages of a rank-based ApEn is that it will be more resistant to outliers, and invariant to monotone transform of data. Another possibility is to discretize raw data such that ApEn’s constructed on discritized data are invariant to monotone transformation and has known asymptotic distributions.

As for the blocklength, instead of having to choose between one and two, one might consider a ‘hybrid’ ApEn, say, a convex combination of these two parameterizations standardized by their covariance.

A proper variance is needed for hypothesis testings based on ApEn. Since asymptotic variances are only available for some special cases, one may consider bootstrap to approximate the distributions of estimated ApEn indices. Due to the fact that ApEn is commonly applied to time series data, bootstrap methods for dependent data, such as stationary bootstrap or block bootstrap, can be useful here.

Lastly, it appears that all above discussions apply to Cross-ApEn, the counterpart of mutual information in information theory.

*Texas A&M University; email: xwu@ag.tamu.edu
Philosophy of Information

Pieter Adriaans

I believe that the main ambition of philosophy of information at this moment in time is to develop a unified theory of Entropy, Information and Computation. Building on work by authors like Li and Vitanyi, Cover and Thomas, Gell-Mann and Lloyd and Bais and Farmer the perspectives for such a theory have in the last decade become much more promising. Three possible contributing theories are Shannon’s Information theory, Gibb’s notion of entropy and Kolmogorov complexity. Of these three I believe that Kolmogorov complexity offers the best starting point for a unified theory for the following reasons:

- The theory of Kolmogorov complexity gives us an objective mathematical measure of entropy for a system, without any compromising epistemological interpretations in terms of beliefs, probabilities or knowledge. A random string x has objective maximum entropy, whatever, we believe or know about it.
- The definition of Kolmogorov complexity in terms of all possible programs, can be interpreted as a definition of entropy quantifying over all possible descriptions of deterministic processes. Any set of mathematical formulas a physicist, biologist or any other scientist might think of to describe processes, in which an entropy-like phenomenon is to occur, is subsumed by the Kolmogorov model of entropy.
- Based on these results, one can give a systematic treatment of the informational aspects of various physical, biological and computational processes.

Using these insights I define the notion of Facticity as a measure for the interestingness of a data set as: the length of the model information for that set under optimal two-part code compression. The notion of facticity is close to other proposals like effective complexity by Gell-Mann and Lloyd and Self-dissimilarity by Wolpert and Mcready. New is the idea to distinguish factic processes that maximize facticity. Factic processes are characterized by model aversion and randomness aversion. They have no sufficient statistic and are abundant in nature: learning, game playing, teaching, evolution, stock prices etc. In the coming years I hope to be working at a better understanding of these phenomena, specifically:

- the description of the evolution processes in computational spaces,
- the nature of various probability distributions that dominate logical, physical, biological and cultural domains,
- the algorithmic approximation of various compression measures,
- the study of cognition and learning as data compression
Using Self-dissimilarity to Quantify Complexity

David Wolpert, NASA

For many systems characterized as “complex” the patterns exhibited on different scales differ markedly from one another. For example the biomass distribution in a human body “looks very different” depending on the scale at which one examines it. Conversely, the patterns at different scales in “simple” systems (e.g., gases, mountains, crystals) vary little from one scale to another. Accordingly, the degrees of self-dissimilarity between the patterns of a system at various scales constitute a complexity “signature” of that system. Here we present a novel quantification of self-dissimilarity. This signature can, if desired, incorporate a novel information-theoretic measure of the distance between probability distributions that we derive here. Whatever distance measure is chosen, our quantification of self-dissimilarity can be measured for many kinds of real-world data. This allows comparisons of the complexity signatures of wholly different kinds of systems (e.g., systems involving information density in a digital computer vs. species densities in a rain-forest vs. capital density in an economy, etc.). Moreover, in contrast to many other suggested complexity measures, evaluating the self-dissimilarity of a system does not require one to already have a model of the system. These facts may allow self-dissimilarity signatures to be used as the underlying observational variables of an eventual overarching theory relating all complex systems. To illustrate self-dissimilarity we present several numerical experiments. In particular, we show that underlying structure of the logistic map is picked out by the self-dissimilarity signature of time series' produced by that map.
Comments on “Using Self-Dissimilarity to Quantify Complexity” by David Wolpert

Douglas Miller
University of Missouri

This paper develops a scalar divergence functional to represent the self-dissimilarity among the characteristics of a system. The characteristics are evaluated under probability models at various scales, and the divergence tools proposed in the paper form an overall index of the degree of self-dissimilarity across all scale categories. Accordingly, this method should provide a more comprehensive approach to representing complexity than other commonly used methods that do not consider multiply scales. In particular, as noted in the paper, this approach should offer advantages relative to transformation-based methods such as Fourier (spectral) or wavelet transforms that only focus on complexity for one characteristic or in one space.

There may be several plausible applications of this approach to quantifying the degree of complexity in social science research, and I will only mention five examples that seem to be natural applications. First, the paper develops the method in the context of spatial-temporal data, and we now use an increasing amount of repeated spatial, longitudinal, and panel data in our research. Second, the examples presented in the paper are reasonably similar to the data analysis problems that arise in the new field of neuro-economics. Third, researchers in marketing and other social science fields are already using complexity-based tools to find patterns in their massive data sets, and the proposed approach should be attractive for these applications due to its potential comparative advantages. Fourth, there is a small but emerging literature in the management field that considers the factors that determine the potential success or failure of strategic alliances formed by two distinct firms. Although much of this research is relatively informal, the authors are asking questions that are closely related to the fundamental problems addressed in this paper --- how do we measure the multiple characteristic of firms, what factors contribute to their overall complexity or self-dissimilarity, and do strategic alliances fail because the firms are too similar (i.e., the relationship does not generate synergies because the partners or firms do not have complementary strengths) or too dissimilar (i.e., the partners have difficulty communicating and are unable to effectively cooperate)? Finally, the approach may provide a simple and feasible way to handle nuisance parameters or to conduct model selection.

There are also some methodological issues that must be addressed before this approach may be widely adopted in social science research. Although the proposed method is not based on extensive models of the structural relationships in the candidate systems, the key disadvantage of the current self-dissimilarity formulation is its heavy reliance on explicit probability models. In contrast, the general direction in econometrics and other quantitative social science fields is to rely on more flexible (e.g., semiparametric or nonparametric) probability models with a bit more focus on the structure of the behavioral relationships. As noted in the paper, the probability arguments in the divergence functionals may be replaced by likelihood values, but we would have to know some additional details before making general applications of this approach in econometric research. For example, is the proposed divergence functional uniformly convergent, perhaps under bootstrap estimators of the likelihood values? If we use these estimators for cluster identification (as in the paper), are the estimators consistent? Can we derive limiting distributions for the self-dissimilarity estimators in order to formally test cluster composition?
Roundtable summary: Information Processing across the Sciences
Robin L. Lumsdaine (moderator)

The conference concluded with a wrap-up of what we had learned over the day and particularly what social scientists could take from our science and philosophy colleagues regarding info-metrics techniques and ideas. Because much of the day had focused on the similarities, overlap, and synergies between various disciplines, to start the discussion off panelists were asked their views on the question of what they viewed as the main differences between info-metrics’ use in the natural sciences versus the social sciences and what advice they might have for use in the social sciences.

The moderator began the discussion by linking some thoughts that had been offered earlier in the day to concepts from the social sciences. For example, Raphy Levine had noted that the concept of probability was subjective which brought to mind the subjective probability literature in economics, as well as implications for marketing and survey response fields. Similarly, Ariel Caticha’s presentation suggested that the treatment of time may differ across fields, with time traditionally being a more unidirectional, linear concept in social science analysis. Essie Maasoumi mentioned the role of conditioning information and how much conditioning was really necessary in setting up a problem. In addition, David Wolpert’s and Pieter Adriaans’ presentations regarding identifying irregular or complex patterns could be useful for the world of risk management and finance, for example, helping policymakers to identify anomalies in trading behavior patterns.

The discussion then turned to a spirited exchange on the difference between the maximum entropy and Bayesian approaches.
My remarks mentioned each of the thought-provoking presented papers, often critically, in order to illustrate the force of the Bayesian argument. I will not refer to any particular presentation here.

My first comment was “focus.” It is crucial to specify the goal of the analysis – an inference, a decision? Then specify what quantities are known and what are unknown, certain or uncertain. This simple step often causes confusion at the outset. In an estimation problem parameters might be fixed but unknown. Data are actually seen and hence known. Uncertainty is quantified coherently by probabilities. Thus a probability distribution for parameters is needed, and a likelihood function to allow data evidence to be accumulated coherently using Bayes Rule. Both specifications are subjective. Coherence describes how probabilities fit together. If there are two events considered, A and B, we might find it convenient to think about P(A), P(B), P(A,B), P(A|B), P(A c), etc. Coherence requires that the system of probabilities make sense: specifying P(A), P(B) and P(A,B) for example determines the full system of probabilities. Quantifications of uncertainty that do not meet these restrictions will inevitably lead to errors. These remarks are made forcefully (and with specifics) by de Finneti, who also shows that coherence can be obtained by primitive assumptions, for example avoiding certain losses in betting. The Bayesian approach is coherent.

Information theory is not alternative or contradictory to the Bayesian approach. Indeed, entropy considerations can lead to insight in the specification of priors or likelihood functions. A device in several presentations was to specify a set of distributions satisfying certain properties, and then choose an element by projecting a uniform distribution to the set using a relative entropy projection. This is sensible, but should not be confused with Bayesian updating. The language of “prior” referring to the uniform and “posterior” referring to the projection is better avoided.

Professor Lumsdaine summarized my remarks succinctly as “Kiefer argues that probabilities are our building blocks and that probabilities are subjective.” Yes.