Information **2011**, *2*, 17-40; doi:10.3390/info2010017

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

Empirical Information Metrics for Prediction Power and Experiment Planning

Christopher Lee [1,2,3]

[1] Department of Chemistry & Biochemistry, University of California, Los Angeles, CA 90095, USA
[2] Department of Computer Science, University of California, Los Angeles, CA 90095, USA
[3] Institute for Genomics & Proteomics, University of California, Los Angeles, CA 90095, USA;
E-Mail: leec@chem.ucla.edu; Tel: 310-825-7374; Fax: 310-206-7286

Received: 8 October 2010; in revised form: 30 November 2010 / Accepted: 21 December 2010 / Published: 11 January 2011

Abstract: In principle, information theory could provide useful metrics for statistical inference. In practice this is impeded by divergent assumptions: Information theory assumes the joint distribution of variables of interest is known, whereas in statistical inference it is hidden and is the goal of inference. To integrate these approaches we note a common theme they share, namely the measurement of *prediction power*. We generalize this concept as an information metric, subject to several requirements: Calculation of the metric must be *objective* or model-free; unbiased; convergent; probabilistically bounded; and low in computational complexity. Unfortunately, widely used model selection metrics such as Maximum Likelihood, the Akaike Information Criterion and Bayesian Information Criterion do not necessarily meet all these requirements. We define four distinct empirical information metrics measured via sampling, with explicit Law of Large Numbers convergence guarantees, which meet these requirements: I_e, the *empirical information*, a measure of average prediction power; I_b, the *overfitting bias information*, which measures selection bias in the modeling procedure; I_p, the *potential information*, which measures the total remaining information in the observations *not* yet discovered by the model; and I_m, the *model information*, which measures the model's extrapolation prediction power. Finally, we show that $I_p + I_e$, $I_p + I_m$, and $I_e - I_m$ are fixed constants for a given observed dataset (i.e. prediction target), independent of the model, and thus represent a fundamental subdivision of the total information contained in the observations. We discuss the application of these metrics to modeling and experiment planning.

Keywords: information metric; inference; sampling; entropy; experiment planning

1. Introduction

1.1. The Need for Information Metrics for Statistical and Scientific Inference

Information theory as formulated by Shannon [1], Kolmogorov and others provides an elegant and general measure of information (or *coupling*) that connects variables. As such, it might be expected to be universally applied in the "Information Age" (see, for example, the many fields to which it is relevant, described in [2]). Identifying and measuring such information connections between variables lies at the heart of statistical inference (infering accurate models from observed data) and more generally of scientific inference (performing experimental observations to infer increasingly accurate models of the universe).

However, information theory and statistical inference are founded on rather different assumptions, which greatly complicate their union. Statistical inference draws a fundamental distinction between *observable variables* (operationally defined measurements with no uncertainty) and *hidden variables* (everything else). It seeks to estimate the likely probability distribution of a hidden variable(s), given a sample of relevant observed variables. Note that from this point of view, probability distributions are themselves *hidden*, in the sense that they can only be *estimated* (with some uncertainty) via inference. For example, individual *values* of an observable are directly observed, but their true *distribution* can only be inferred from a sample of many such observations.

Traditional information theory, by contrast, assumes as a starting point that the joint probability distribution $p(X, Y, Z...)$ of all variables of interest is completely known, as a prerequisite for beginning any calculations. The basic tools of information theory – entropy, relative entropy, and mutual information – are undefined unless one has the complete joint probability distribution $p(X, Y, Z...)$ in hand. Unfortunately, in statistical inference problems this joint distribution is unknown, and precisely what we are trying to infer.

Thus, while "marrying" information theory and statistical inference is by no means impossible, it requires clear definitions that resolve these basic mismatches in assumptions. In this paper we begin from a common theme that is important to both areas, namely the concept of *prediction power*, *i.e.*, a model's ability to accurately predict values of the observable variable(s) that it seeks to model. Prediction power metrics have long played a central role in statistical inference. Fisher formulated prediction power as simply the total likelihood of the observations given the model, and developed Maximum Likelihood estimators, based on seeking the specific model that maximizes this quantity. This concept remains central to more recent metrics such as the Akaike Information Criterion (AIC) [3], and Bayesian Information Criterion (BIC) [4], which add "corrections" based on the number of model parameters being fitted.

In this paper we define a set of statistical inference metrics that constitute statistical inference proxies for the fundamental metrics of information theory (such as mutual information, entropy and relative entropy). We show that they are vitally useful for statistical inference (for precisely the same properties that make them useful in information theory), and highlight how they differ from standard statistical inference metrics such as Maximum Likelihood, AIC and BIC. We present a series of metrics that address distinct aspects of statistical inference:

- *prediction power*, as it is ordinarily defined, as the likelihood of future observations (e.g., "test data") under a given set of conditions that we have already observed ("training data").
- *bias*: A measure of any systematic difference in the model's prediction power on future observations *vs.* on its original training data.
- *completeness*: We define a modeling process as "complete" when no further improvements in prediction power are possible (by further varying the model). Thus a completeness metric measures how far we are from obtaining the best possible model.
- *extrapolation prediction power*:We will introduce a measure of how much the model's prediction power exceeds the prediction power of our existing observation density, when tested on future observations. If this value is zero (or negative) one might reasonably ask to what extent its results can truly be called a "prediction", but instead are only a summary (or "interpolation") of our existing observation data.

To clarify the challenges that such metrics must solve, we wish to highlight several characteristics they must possess:

- *objective* or *model-free*: One important criterion for such a metric is whether it is model-free; that is, whether or not the calculation of the metric itself involves a process that is equivalent to modeling. If it does, the metric can only be considered to yield a "subjective" evaluation – how well one model fits to the expectations of another model. By contrast, a model-free metric aims to provide an objective measure of how well a model fits the empirical observations. While this criterion may seem very simple to achieve, it poses several challenges, which this paper will seek to clarify.
- *unbiased*: Like any estimator calculated from a finite sample, these metrics are expected to suffer from *sampling errors*, but they must be mathematically proven to be free from *systematic errors*. Such errors are an important source of overfitting problems, and it is important to understand how to exclude them by design.
- *convergent*: These metrics must provide explicit Law of Large Numbers proofs that they converge to the "true value" in the limit of large sample size. The assumption of convergence is implicit in the use of many methods (such as Maximum Likelihood), but unfortunately the strict requirements of the Law of Large Numbers are sometimes violated, breaking the convergence guarantee and resulting in serious errors. To prevent this, a metric must explicitly show that it meets the requirements of the Law of Large Numbers.
- *bounded*: These metrics must provide probabilistic bounds that measure the level of uncertainty about their true value, based on the limitations of the available evidence.

- *low computational complexity*: Ideally, the computational complexity for computing a metric should be $O(N \log N)$ or better, where N is the number of sample observations.

In this paper we define a set of metrics obeying these requirements, which we shall refer to as *empirical information metrics*. As a Supplement, we also provide a tutorial that shows how to calculate these metrics using **darwin**, an easy-to-use open source software package in Python, available at https://github.com/cjlee112/darwin.

2. Empirical Information

2.1. Standard Prediction Power Metrics

Fisher defined the prediction power of a model Ψ for an observable variable X in terms of the total likelihood of a sample of independent and identically distributed (I.I.D.) draws $X_1, X_2, ...X_n$

$$p(X_1, X_2, ...X_n|\Psi) = \prod_{i=1}^{n} \Psi(X_i) = \exp\left(\sum_{i=1}^{n} \log \Psi(X_i)\right) = \exp(n\overline{L})$$

where we adopt the convention $\Psi(X) \equiv p(X|\Psi)$ as a shorthand for the probability of an observation given a model, and define the log-likelihood $L = \log \Psi(X)$. We follow the standard notation $\overline{L}$ to indicate its sample mean. Note that we will sometimes write $L(\Psi)$ to emphasize that L is a function of the specific model we are computing.

Fisher's Maximum Likelihood method seeks the model that maximizes the total likelihood or, equivalently, the sample average log-likelihood $\overline{L}$. Similarly, minimizing the Akaike Information Criterion (AIC) [3]

$$AIC = 2k - 2\log p(x_1, x_2, ...x_n|\Psi) = 2k - 2n\overline{L}$$

or the Bayesian Information Criterion (BIC) [4]

$$BIC = k \log n - 2n\overline{L}$$

again seeks to maximize the prediction power $\overline{L}$ while explicitly correcting for model complexity expressed as k, the number of free parameters in the model Ψ.

Vapnik-Chervonenkis theory also supplies a correction factor that penalizes model complexity for classifier problems [6]. For example, consider the simplest case of a binary classifier that predicts the class of each data point with a confidence factor C (by assigning that class a likelihood of $1 - \frac{1}{C}$, and the other class a likelihood of $\frac{1}{C}$). In this case the classification error probability on the training data, R_{train}, converges for large C to $R_{train} \rightarrow -\overline{L}/\log C$, and structural risk minimization indicates choosing the model that minimizes the upper bound of the classification error probability:

$$R_{VC} = \sqrt{\frac{h(1 + \log \frac{2n}{h}) - \log \frac{\eta}{4}}{n}} - \frac{\overline{L}}{\log C}$$

where h is the Vapnik-Chervonenkis (VC) dimension of the model (a measure of model complexity), and η is the desired level of confidence for the probabilistic bound.

2.2. Prediction Power and the Law of Large Numbers

These metrics are best understood by highlighting the critical role that the Law of Large Numbers plays in inference metrics. Say we want to find a model Ψ that maximizes the total likelihood of many draws of X, or equivalently the expectation value of the log-likelihood, which depends on the true distribution $\Omega(X)$:

$$E(L) \equiv \sum_X \Omega(X) \log \Psi(X)$$

where the summation is over all possible values of X (for a continuous variable the summation is replaced by an integral).

Since we do not know the true distribution $\Omega(X)$ we cannot use this definition directly. However, we can apply the Law of Large Numbers (LLN) to the log-likelihood of a sample of observations, whose sample average must converge

$$\overline{L} = \frac{1}{n}\sum_{i=1}^{n} L_i = \frac{1}{n}\sum_{i=1}^{n} \log \Psi(X_i) \overset{LLN}{\longrightarrow} E(L)$$

as $n \to \infty$, if the sample values L_i are conditionally independent given Ω and identically distributed as L, and the variance $Var(L)$ is finite (the LLN can also be extended to the case of exchangeable observations [5]). Specifically, the Law of Large Numbers guarantees a probabilistic bound on the sample estimator's deviation from the expectation value:

$$p\left(|\overline{L} - E(L)| \geq \delta\right) \leq \frac{Var(L)}{n\delta^2}$$

So we obtain a lower bound estimate for L at confidence level $1-\epsilon$ of

$$L_\epsilon = \overline{L} - \sqrt{\frac{Var(L)}{n\epsilon}}$$

Note that to actually compute this lower bound, we must also use our sample to estimate the variance, which adds another source of error. In practice this is usually not a problem, except for pathological cases (e.g., $Var(L) \to \infty$). For example, to calculate a 95% confidence lower bound:

$$L_{0.05} = \overline{L} - \sqrt{\frac{\overline{Var(L)}}{n(0.05)}}$$

where we have used the shorthand notation $\overline{Var(L)} = \overline{(L-\overline{L})^2}$ to denote the sample estimator of the variance. Note that since the Law of Large Numbers is a general result (*i.e.*, it holds over all possible distributions), it does not necessarily represent the best confidence interval that one can obtain for a specific case. Other methods for computing a confidence interval such as resampling [7], can usually improve on (*i.e.*, increase) this lower bound, but we will not explore such implementation details in this paper.

Since the X_i are indeed conditionally independent given Ω and identically distributed as X, we expect for large sample size n to be able to use $\overline{L}$ as a proxy for $E(L)$. In that case maximizing $\overline{L}$ also maximizes

$E(L)$, which it is convenient to separate into one term dependent only on Ω and another term dependent on Ψ:

$$E(L(\Psi)) = \sum_X \Omega(X)\left(\log\frac{\Psi(X)}{\Omega(X)} + \log\Omega(X)\right) = \sum_X \Omega(X)\left(-\log\frac{\Omega(X)}{\Psi(X)} + \log\Omega(X)\right)$$
$$= -D(\Omega||\Psi) - H(\Omega(X))$$

where $D(\Omega||\Psi)$ is the relative entropy of model Ψ relative to the true distribution Ω, and $H(\Omega(X))$ is the entropy of the true distribution Ω. Since the right hand term is constant with respect to Ψ, this expression is maximized when $D(\Omega||\Psi)$ is minimized, which occurs iff $\Psi(X) = \Omega(X)$ for all values of X. This guarantees that choosing the model Ψ^* that maximizes $E(L)$ will indeed identify the correct model $\Psi^*(X) = \Omega(X)$.

2.3. *The Problem of Selection Bias*

Unfortunately, there is a catch. This guarantee can only be extended to maximization of the sample log-likelihood $\overline{L}$, if the L_i are identically distributed as L. All of these metrics ($\overline{L}$, AIC, BIC) were designed for use with *model selection*; that is, we compute the metric for each of a large set of models, then select the model that maximizes the likelihood (or minimizes the AIC or BIC). And the very nature of model selection introduces bias into the sample likelihoods [8]. Briefly, if the model Ψ was chosen specifically to *maximize* the values L_i, we *cannot* assume that the L_i are identically distributed as L. Indeed, we expect that the L_i will be biased to higher values than L in general. Therefore the Law of Large Numbers convergence guarantee collapses, and we cannot prove that model selection using $\overline{L}$ will yield the true distribution Ω. Vapnik-Chervonenkis theory seeks to protect against this bias by deriving an upper bound on the possible error due to selection bias [6], based on the model's VC dimension.

First, let's examine this problem from an empirical point of view, by simply defining a metric for measuring the bias. We define a *test data criterion*:

- a set of sample values $X'_1, X'_2, ...X'_m$ are valid *test data* for a model Φ predicting an observable X if the X'_i are exchangeable, identically distributed as X, and conditionally independent of Φ given the true distribution Ω, *i.e.*, $p(X'_i, \Phi|\Omega) = p(X'_i|\Omega)p(\Phi|\Omega)$. Equivalently, Φ contains no information about the X'_i except via their shared dependence on the hidden distribution Ω. Note that for any model Φ generated by model selection, its *training data* do not meet this requirement, since Φ is *not* conditionally independent of the training data given Ω.

We desire an estimator for $\overline{L} - E(L)$. Since the X'_i are identically distributed as X and conditionally independent of Φ given Ω, the $\log\Phi(X'_i)$ are identically distributed as $\log\Phi(X)$ *i.e.*, L. So by the Law of Large Numbers we can define an *overfitting bias information* metric

$$\overline{I_b} = \overline{L} - \overline{L_e} \xrightarrow{LLN} \overline{L} - E(L)$$

as $m \to \infty$, where $\overline{L_e}$ is the sample average of the $\log\Phi(X'_i)$ test data log-likelihoods. We will refer to $\overline{L_e}$ as the *empirical log-likelihood*. Note that whereas Vapnik-Chervonenkis theory provides an *upper bound* on the bias errors for an entire class of models (*i.e.*, all models with the same VC dimension), I_b measures the *actual error* due to a specific model's selection bias.

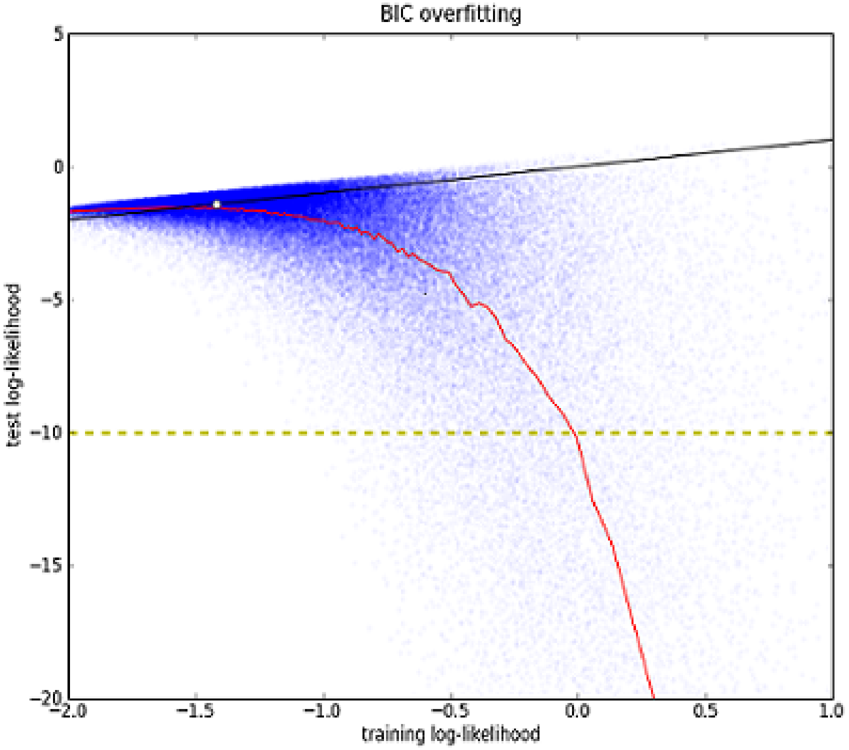
I_b has the corresponding lower bound estimator (under the simplifying assumption that the sample sizes for L and L_e are the same $(m = n)$):

$$I_{b,\epsilon} = \overline{I_b} - \sqrt{\frac{\overline{Var(L - L_e)}}{m\epsilon}}$$

If the model selection procedure has introduced no bias, $I_b \approx 0$.

2.4. Example: The BIC Optimal Model for a Small Sample from a Normal Distribution

Figure 1. Overfitting analysis of BIC models on a small sample from a normal distribution. For each data point, a sample of three observations was drawn randomly from a unit normal distribution. The BIC-optimal model was fit to these observations and used to compute the training *vs.* test log-likelihoods $\overline{L}$ *vs.* $\overline{L_e}$, the latter calculated on an additional test sample of three observations drawn from the same unit normal. To generate the scatter plot, this process was performed a total of $N = 100000$ times. The mean value of $\overline{L_e}$ for successive windows of 1000 observations sorted from left to right is plotted in red. The zero-bias line is shown in black ($\overline{L} = \overline{L_e}$). Thus, the overfitting bias information $\overline{I_b}$ is given at any position on the graph by the vertical distance between the black and red lines. The white circle indicates the true expectation log-likelihood for the unit normal distribution. The dotted line marks the mean value of $\overline{L_e}$ averaged over all 100,000 data points. Note that this figure shows only a portion of the full distribution, which has a long tail extending to large negative values of $\overline{L_e}$.



The BIC adds a correction term $k \log n$ to the total log-likelihood, which penalizes against models with larger numbers of parameters. Note that this correction is designed specifically to protect against overfitting. This correction is referred to as the Bayesian Information Criterion because it is based on choosing the model with maximum Bayesian posterior probability, and by this criterion is provably optimal for the exponential family of models [4].

However, several caveats about such corrections should be understood:

- a given correction addresses a particular kind of overfitting, for example, for the AIC and BIC, excessive number of model parameters k.
- a given correction is based on specific assumptions about the model, and may not behave as expected under other conditions;
- Such corrections do *not* guarantee that the model they select will be optimal, or even unbiased.

As an example, *Figure 1: Overfitting analysis of BIC models on a small sample from a normal distribution* shows the distribution of $\overline{L}$ *vs.* $\overline{L_e}$ for BIC-optimal models generated using a sample of three observations drawn from a unit normal distribution. (Note that in this case BIC-optimality is just equivalent to AIC-optimality and Maximum Likelihood, since the set of all possible normal models share the same value of $k = 2$). This simple example illustrates several points:

- A large fraction of the models strongly overfit the observations as indicated by a large deviation from the $\overline{L} = \overline{L_e}$ diagonal.
- $\overline{L}$ and $\overline{L_e}$ are strongly and non-linearly anti-correlated. That is, the better the apparent fit to the training data, the worse the actual fit to the test data.

2.5. *The Empirical Information Metric*

Based on these considerations, we use the unbiased estimator L_e to define the *empirical information*, a signed measure of prediction power relative to the uninformative distribution $p(X)$:

$$\overline{I_e(\Psi)} = \overline{L_e(\Psi)} - \overline{L_e(p)} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Psi(X_i)}{p(X_i)}$$

The empirical information estimates the improvement in the accuracy of a model $\Psi(X)$ in predicting the test observations. For observable variables X whose uninformative distribution is simply a constant density, $\overline{I_e(\Psi)}$ differs from $\overline{L_e(\Psi)}$ by simply a constant ($\log R$, where R is the size of the range of X). In such cases the lower bound estimator for I_e differs from that of L_e only by this constant:

$$I_{e,\epsilon} = \overline{I_e} - \sqrt{\frac{\overline{Var(L_e)}}{n\epsilon}} = \overline{L_e} - \sqrt{\frac{\overline{Var(L_e)}}{n\epsilon}} + \log R$$

It is important to note a few aspects of the empirical information that arise from the above considerations:

- Note that I_e can be negative, if the model's prediction power is even worse than that of the uninformative distribution.

- Whereas most metrics for model selection such as the AIC and BIC contain correction terms dependent on the model complexity k (or VC dimension h), I_e needs no such corrections because it is unbiased *by definition*. Excessive model complexity will not increase I_e but instead will reduce it. I_e contains no bias and therefore needs no correction. In this sense, it follows a similar approach as cross-validation [9].
- Note that we do not need to incorporate the sample size directly into the metric definition (as in the case of the BIC [4], Vapnik-Chervonenkis upper-bound error R_{VC} [6], and "small-sample corrected" versions of the AIC such as the AICc [10]). Instead, the effect of sample size emerges naturally from the Law of Large Numbers lower bound estimator for our empirical information metrics (e.g., $L_{e,\epsilon}, I_{b,\epsilon}, I_{e,\epsilon}$). Fundamentally, the importance of sample size is simply the uncertainty due to sampling error, and the Law of Large Numbers probabilistic bound captures this in a general way.

2.6. *Empirical Information as A Sampleable Form of Mutual Information*

Consider the following "mutual information sampling problem":

- draw a specific inference problem (hidden distribution $\Omega(X)$) from some class of real-world problems (e.g., for weight distributions of different animal species, this step would mean randomly choosing one particular animal species);
- draw training data $\vec{X}^t$ and test data X from $\Omega(X)$;
- find a way to estimate the mutual information $I(\vec{X}^t; X)$ on the basis of this single case (single instance of Ω).

The standard definition of mutual information $I(\vec{X}^t; X) = E\left(\log \frac{p(\vec{X}^t, X)}{p(\vec{X}^t)p(X)}\right)$ does not enable such a calculation. Even if we draw many pairs $\vec{X}^t, X$ to estimate this value, we will just get a value of zero, because $\vec{X}^t, X$ are conditionally independent given Ω. The mutual information $I(\vec{X}^t; X)$ is defined only over the *complete* joint distribution $p(\Omega, \vec{X}^t, X)$; it does not appear meaningful to talk about calculating it from a single instance of Ω.

By contrast with mutual information, we *do* calculate empirical information for a specific value of Ω, *i.e.*, we use it to measure the prediction power of our model Ψ on observations emitted by that specific value of Ω. It is therefore interesting to investigate the relationship of the empirical information *vs.* the mutual information. We follow the usual information theory approach of taking its expectation value over the complete joint distribution:

$$E(I_e(\Psi)) = E(L_e(\Psi)) - E(L_e(p)) = E(L_e(\Psi)) - \sum_X p(X) \log p(X) = E(L_e(\Psi)) + H(X)$$

assuming that the uninformative distribution $p(X)$ used in the denominator of I_e matches the true marginal distribution of X. Focusing on the remaining expectation log-likelihood term:

$$E(L_e(\Psi)) = \sum_\Omega \sum_{\vec{X}^t} \sum_X p(X, \vec{X}^t, \Omega) \log \Psi(X|\vec{X}^t)$$

where we take the expectation value over all possible values of the observable X, all possible values of the hidden variable Ω, and all possible training data sets $\vec{X}^t$ of size t. Note that we write the model as $\Psi(X|\vec{X}^t)$ to explicitly emphasize its dependence on a set of training data $\vec{X}^t$. Since Ω does not appear in the log term we can eliminate it:

$$= \sum_{\vec{X}^t} \sum_X p(X, \vec{X}^t) \log \Psi(X|\vec{X}^t)$$

$$= -\sum_{\vec{X}^t} p(\vec{X}^t) \sum_X p(X|\vec{X}^t) \log \frac{p(X|\vec{X}^t)}{\Psi(X|\vec{X}^t)} + \sum_{\vec{X}^t} \sum_X p(X, \vec{X}^t) \log p(X|\vec{X}^t)$$
$$= -E_{\vec{X}^t}(D(p(X|\vec{X}^t)||\Psi(X|\vec{X}^t))) - H(X|\vec{X}^t)$$

where the first term is a relative entropy of the model *vs.* the true conditional probability, and the second term is the conditional entropy of the observable *vs.* the training data. Therefore the expectation value of the empirical information is just:

$$E(I_e(\Psi)) = H(X) - H(X|\vec{X}^t) - E_{\vec{X}^t}(D(p(X|\vec{X}^t)||\Psi(X|\vec{X}^t)))$$
$$= I(X; \vec{X}^t) - E_{\vec{X}^t}(D(p(X|\vec{X}^t)||\Psi(X|\vec{X}^t)))$$

where $I(X; \vec{X}^t)$ is the mutual information between the training data and the observable. Now consider the following sampling protocol:

- for one specific inference problem (hidden value of Ω), we draw a training dataset $\vec{X}^t$, use it to train a model $\Psi(X|\vec{X}^t)$, and measure the empirical information $\overline{I_e(\Psi)}$ on a set of test data $\vec{X}^n$ drawn from the same distribution.
- We repeat this procedure for multiple inference problems $\Omega_{(1)}, \Omega_{(2)}, ..., \Omega_{(m)}$, and take the average of their empirical information values $\frac{1}{m}\sum \overline{I_e} \overset{LLN}{\longrightarrow} E(I_e(\Psi))$.

If the model $\Psi(X|\vec{X}^t)$ approximates the true conditional distribution $p(X|\vec{X}^t)$ more and more closely, the relative entropy term $D(p(X|\vec{X}^t)||\Psi(X|\vec{X}^t))$ will vanish, and we expect the average of the empirical information values to converge simply to $I(X; \vec{X}^t)$. Under these conditions, the empirical information becomes a "sampleable form" of the mutual information. Note that the mutual information itself does *not* have this property; as shown above, the mutual information cannot be computed "piecewise" for individual instances of Ω and then averaged. By contrast, if we compute the empirical information for each inference problem, and then take the average, it will converge to the mutual information.

3. The Problem of Convergence

If we wish to maximize prediction power, our ultimate goal must be convergence, namely that our model will converge to the true, hidden distribution Ω. So we must ask the obvious question, how do we know when we're done? Two basic strategies present themselves:

- *self-consistency tests*: We can use our model as a reference to test whether the observations exactly match its expectations, as must be true if $\Psi \to \Omega$.

- *convergence distance metric*: If we knew the value of the absolute maximum prediction power $L(\Omega)$ possible for our target observable X, we could define a distance metric $\delta = L(\Omega) - L(\Psi)$, which measures how "far" our current model is from convergence, in terms of its relative prediction power.

We will define empirical information metrics for both these approaches.

3.1. The Inference "Halting Problem"

As an example of the need for a convergence metric, we consider the process of Bayesian inference in modeling scientific data. In scientific research, we cannot easily restrict the set of possible models *a priori* either to closed-form analytic solutions or to finite sets of models that we can fully compute in practical amounts of CPU time. That is, the set of all possible models of the universe is not strictly bounded, and generally can be reduced only by calculating likelihoods for different terms of this set *vs.* experimental observations.

What is the computational complexity for Bayesian inference to find the correct term Ω or any term within some distance δ of it? We can view this as a form of the Halting Problem, in the sense that it requires a metric that indicates when it has found a term that is less than δ distance from Ω, at which point the algorithm halts. Unfortunately, the standard form of Bayes' Law

$$p(\Psi|\vec{X}) = \frac{p(\vec{X}|\Psi)p(\Psi)}{\sum_\Psi p(\vec{X}|\Psi)p(\Psi)}$$

offers no evident shortcuts: Even if we had calculated all but one last term of the summation, we still would not know whether our best model so far is actually the best model, or even whether it is within distance δ of the best model. In the absence of a halting test, this implies that its computational complexity must simply be that of exhaustive enumeration. This is a serious problem, especially given that the set of all possible models may for scientific inference problems be infinite.

In real-world practice this "halting problem" often grows into an even worse problem of "model misspecification" [11]. That is, Bayesian computational methods typically lack a mechanism for generating *all* possible models even in theory. Instead they are limited to assuming a specific mathematical form for the model. Unless by good fortune the true distribution exactly fits this mathematical form, the computation will simply exclude it. Therefore, a reliable convergence metric becomes essential as an external indicator for whether the computational model is "misspecified" in this way. It should be noted that this is *not* addressed by asking whether a given Bayesian modeling process has "converged" in the sense of a Markov Chain Monte Carlo sampling process converging to its stationary distribution [12]. Any such process is still restricted by its assumptions of a specific mathematical form for the model; there is no guarantee that this will contain the correct answer.

3.2. Potential Information

We define I_∞ as the total information content obtainable from a set of observations by considering the infinite set of all possible models. By analogy to the classical physics division of kinetic *vs.* potential energy components, we divide this into one part representing the model terms we've actually calculated

(I_e, the empirical information), and a second part for the remaining *uncomputed terms*, which we define as I_p, the **potential information**:

$$I_\infty = I_e + I_p$$

I_p therefore represents the maximum amount of information theoretically attainable by computing more terms of the infinite set. Assuming that the true, hidden likelihood is $\Omega(X)$ and that our current model (after considering all terms calculated so far) is $\Psi(X)$, then

$$I_p = I_\infty - I_e = \sum_{\forall x} \Omega(x) \log \frac{\Omega(x)}{p(X)} - \sum_{\forall x} \Omega(x) \log \frac{\Psi(x)}{p(X)}$$

where $p(X)$ is the uninformative reference distribution, which cancels, yielding

$$I_p = \sum_{\forall x} \Omega(x) \log \Omega(x) - \sum_{\forall x} \Omega(x) \log \Psi(x) = -H(\Omega(X)) - E(L)$$

We can therefore solve the Inference Halting Problem by deriving an empirical I_p estimator (with a Law of Large Numbers convergence guarantee) that can be calculated *without computing any more terms of the infinite model set*. This is surprisingly straightforward. The right-hand term can be estimated directly by $-\overline{L_e}$ (the empirical log-likelihood). The left-hand term $-H(\Omega(X))$ is simply the negative entropy of the observable. We evidently need an empirical estimator of the entropy, and specifically of the density $\Omega(X)$.

This density estimation problem poses one conceptual problem that requires clarification. Since the ultimate purpose of the potential information calculation is to catch possible errors in modeling, no part of its calculation (such as the empirical entropy calculation) should itself be equivalent to a form of modeling. If we used such a form of modeling to compute the empirical entropy, that would introduce a strongly subjective element, *i.e.*, simply comparing one model (Ψ) versus another (the model used for estimating H_e). To obtain an *objective* I_p metric, the empirical entropy calculation should be *model-free*. It should be a purely empirical procedure with a Law of Large Numbers convergence guarantee for large sample size $n \to \infty$.

3.3. The Empirical Entropy

For the case where the observable X is restricted to a set of discrete values, we define an indicator label $\kappa_x(X)$ which equals 1 if X equals a desired value x, otherwise zero. Then by the Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_x(X_i) \overset{LLN}{\longrightarrow} E(\kappa_x(X)) = \Omega(X = x)(1) + \Omega(X \neq x)(0) = \Omega(x)$$

The empirical entropy estimator follows directly in this case: for $n \to \infty$,

$$\overline{H_e} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{n}\sum_{j=1}^{n} \kappa_{X_i}(X_j)\right) \overset{LLN}{\longrightarrow} H(\Omega(X))$$

For the continuous case, we need an empirical probability density estimator $P_e(X)$. To obtain this we define an indicator function $\kappa_x(X)$ which equals 1 if $X \leq x$, otherwise zero. Then

$$\frac{1}{n}\sum_{i=1}^{n} \kappa_x(X_i) \overset{LLN}{\longrightarrow} E(\kappa_x(X)) = \int_{-\infty}^{\infty} \Omega(X)\kappa_x(X)dX = \int_{-\infty}^{x} \Omega(X)dX$$

i.e., the cumulative density function $c.d.f.(X)$. Therefore we define

$$\overline{H_e} = -\frac{1}{n}\sum_{j=1}^{n} \log P_e(X_j) = -\frac{1}{n}\sum_{j=1}^{n} \log \frac{\sum_{i=1}^{n} \kappa_{X_j+\delta x/2}(X_i) - \sum_{i=1}^{n} \kappa_{X_j-\delta x/2}(X_i)}{n\delta x}$$

$$\xrightarrow{LLN} -E\left(\log \frac{\sum_{i=1}^{n} \kappa_{X+\delta x/2}(X_i) - \sum_{i=1}^{n} \kappa_{X-\delta x/2}(X_i)}{n\delta x}\right)$$

$$\to -E\left(\log \frac{c.d.f.(X+\delta x/2) - c.d.f.(X-\delta x/2)}{\delta x}\right)$$

By construction we choose $\delta x \propto 1/n \to 0$ as $n \to \infty$. Then by the Fundamental Theorem of Calculus,

$$\overline{H_e} \xrightarrow{LLN} -E(\log \Omega(X)) = -\int_{-\infty}^{\infty} \Omega(X) \log \Omega(X) dX = H(\Omega(X))$$

For example, we can construct $\delta x \propto 1/n$ as follows: For each sample point X_j, find its m nearest neighbors (sample points), where m is a relatively small constant. Then set

$$\delta x = |X_{j:m} - X_j| + |X_{j:m-1} - X_j|$$

where we use the notation $X_{j:m}$ to mean the "m -th nearest neighbor of point X_j" . Note that the interval $[X_j - \delta x/2, X_j + \delta x/2]$ contains $m-1$ sample points (not including X_j itself, to avoid the inherent bias that would introduce; this in turn requires replacing the n in the log-denominator with $n-1$). This implementation of the $\overline{H_e}$ calculation is simply:

$$\overline{H_e} = -\frac{1}{n}\sum_{j=1}^{n} \log \frac{m-1}{(n-1)(|X_{j:m} - X_j| + |X_{j:m-1} - X_j|)}$$

There are of course many possible empirical density estimation implementations that could be used; we offer this implementation solely as an illustrative example. This implementation also generalizes to multidimensional data, and thus can be used to estimate mutual information [13,14].

Of course, the empirical entropy has the usual lower bound estimator from the Law of Large Numbers

$$H_{e,\epsilon} = \overline{H_e} - \sqrt{\frac{\overline{Var(\log P_e)}}{n\epsilon}}$$

3.4. Potential Information Estimators

This gives us mean and lower bounds estimators for the potential information

$$\overline{I_p} = -\overline{H_e} - \overline{L_e}$$

$$I_{p,\epsilon} = \overline{I_p} - \sqrt{\frac{\overline{Var(\log P_e - L_e)}}{n\epsilon}}$$

where the variance is computed from P_e and L_e pairs calculated from the same sample of observations.

Note that since the potential information is computed in "observation space" instead of "model space", the computational complexity of its calculation depends primarily on the observation sample size. This can be very efficient. First of all, the calculation divides into two parts that can be done separately; since the empirical entropy has no dependence on the model Ψ, it need only be calculated once and can then used for computing I_p for many different models. Second, the empirical entropy calculation can have low computational complexity. For the simple implementation outlined above, it is simply $O(mn)$ (where m is a small constant for the nearest-neighbor density calculation; this assumes the observations are already sorted in order. If not, an additional $O(n \log n)$ step is required to sort them). For high dimensional data, the computational complexity scales as $O(n^2)$, due to the need to calculate pairwise distances. Of course, the details of the computational complexity will vary depending on what empirical entropy implementation is used.

3.5. Convergence to the Kullback-Leibler Distance

In the limit of large sample size, the potential information converges to

$$\overline{I_p} \overset{LLN}{\longrightarrow} E(\log \Omega(X) - \log \Psi(X)) = D(\Omega || \Psi)$$

which is simply the relative entropy (Kullback-Leibler divergence [15]) of the true distribution *vs.* the model. (It should be emphasized that computing the Kullback-Leibler divergence directly requires knowing the true distribution, which of course in any inference problem is unknown).

We may thus consider the potential information to represent a distance estimator from the true distribution Ω. Specifically, it estimates the difference in prediction power of our current model *vs.* that of the true distribution. Thus it solves the Inference Halting metric problem; if we are searching for a model with prediction power within distance δ of the maximum, we simply halt when

$$\overline{I_p} + \sqrt{\frac{\overline{Var(\log P_e - L_e)}}{n\epsilon}} \leq \delta$$

at whatever level of confidence $1 - \epsilon$ we desire.

The Akaike Information Criterion (AIC) [3] and related information metrics [16] are often referred to as representing the Kullback-Leibler (KL) divergence of the true distribution *vs.* the model [17]. So it is logical to ask how the potential information differs from these well-known metrics. The AIC and related metrics were designed for *model selection* problems, in which the observable (characterized by the true distribution Ω) is treated as a fixed constant, and the model is varied in search of the best fit. As shown in part **A** of *Figure 2: Comparing AIC and Potential Information to the Theoretical Kullback-Leibler Divergence*, the AIC does indeed correlate directly with the KL divergence $D(\Omega||\Psi)$ under this assumption (holding the true distribution fixed as a constant). Specifically, for a sample of exchangeable observations $\vec{X}^n$,

$$AIC = 2k - 2\log \Psi(\vec{X}^n) = 2k - 2n\overline{L(\Psi)}$$

So as $n \to \infty$,

$$\frac{1}{2n} AIC \overset{LLN}{\longrightarrow} -E(L(\Psi))$$

Thus the AIC converges to the negative log-likelihood, whereas the KL divergence $D(\Omega||\Psi) = -H(\Omega(X)) - E(L(\Psi))$ also contains an entropy term $-H(\Omega(X))$. However, if the true distribution $\Omega(X)$ is held fixed, then the AIC differs from the KL divergence only by a constant. So for comparing two different models Ψ_1, Ψ_2, the difference in their AIC values converges to

$$\frac{1}{2n}(AIC(\Psi_2) - AIC(\Psi_1)) \overset{LLN}{\longrightarrow} D(\Omega||\Psi_2) - D(\Omega||\Psi_1)$$

This is why the AIC and related likelihood metrics are often treated as a proxy for the KL divergence in model selection.

However, if the true distribution Ω is *not* treated as a fixed constant, and instead is allowed to vary, this simple relationship breaks. In that case, the AIC no longer correlates with the KL divergence (Figure 2B). By contrast, the potential information metric $\overline{I_p}$ correlates with the KL divergence under *all* conditions (Figure 2C). The main difference between the potential information and the AIC is simply the empirical entropy term, which is included in the potential information metric but missing from the AIC:

$$\overline{I_p} - \frac{1}{2n}AIC = -\overline{H_e} - \frac{k}{n}$$

Thus, the potential information metric (and consequently, the empirical entropy term) is essential for any problem where

- we need an estimate of the *absolute* value of the Kullback-Leibler divergence, rather than simply comparing its *relative* value for two models;
- or we need to consider possible variation between *different* true distributions Ω (or equivalently, different observable variables X). For example, in *experiment planning* problems, we consider different possible experiments (different observable variables) in order to estimate how much information they are likely to yield [18].

3.6. *Unbiased Empirical Posteriors*

Standard Bayesian inference can grossly overestimate the posterior probability of a model term, because the sum of calculated terms is biased to underestimate the total $p(X)$ summed over the complete infinite series. The empirical entropy provides a resolution to this problem. By the Asymptotic Equipartition theorem [1], for a sample $\vec{X}^N = \{X_1, X_2, ...X_N\}$ of exchangeable observations of size N

$$\frac{1}{N}\sum_{i=1}^{N}\log p(X_i) \overset{LLN}{\longrightarrow} E(\log p(X)) = \sum_{\forall X}\Omega(X)\log\Omega(X) = -H(\Omega(X))$$

and thus we can therefore estimate $p(\vec{X}^N)$ via

$$p(\vec{X}^N) = \prod_{i=1}^{N} p(X_i) \overset{LLN}{\longrightarrow} \exp(-NH_e)$$

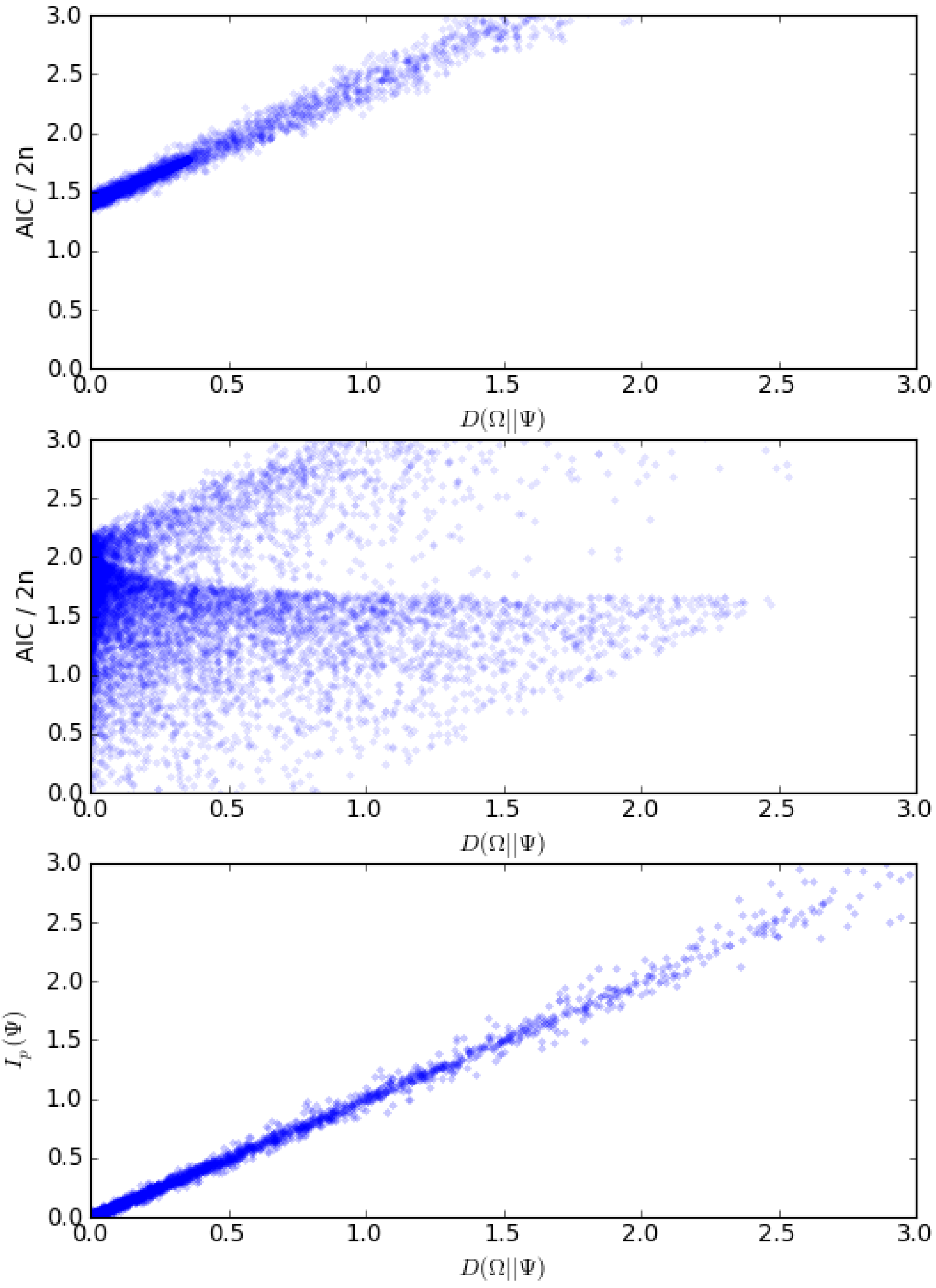
This provides an unbiased estimator of the posterior probability of a model term θ

$$p_e(\theta|\vec{X}^N) = \frac{p(\vec{X}^N|\theta)p(\theta)}{\exp(-NH_e)}$$

We designate this the "empirical posterior" probability of model term θ, with confidence interval:

$$p\left(p_e(\theta|\vec{X}^N)\exp(-N\delta) \leq p(\theta|\vec{X}^N) \leq p_e(\theta|\vec{X}^N)\exp(N\delta)\right) \geq 1 - \frac{4Var(\log P_e)}{N\delta^2}$$

Figure 2. Comparing AIC and Potential Information to the Theoretical Kullback-Leibler Divergence. **A.** Comparison of AIC values *vs.* Kullback-Leibler divergence for a sample of 10,000 different models, with the true distribution fixed to the unit normal distribution $N(0, 1)$. Each model was a normal distribution $N(0, \tau^2)$ *where the standard deviation* τ was drawn uniformly on the interva $(0.1, 2)$. For each model, the AIC was calculated using $n = 1000$ observations. **B.** The same comparison, with a variable true distribution $\Omega = N(0, \sigma^2)$ with standard deviation $\sigma \in (0.1, 2)$. Note the AIC no longer correlates with the Kullback-Leibler divergence. **C.** The same comparison as in **B**, except using the potential information metric. Note that it closely matches the theoretical Kullback-Leibler divergence $D(N(0, \sigma^2)||N(0, \tau^2)) = \log \frac{\tau}{\sigma} + \frac{\sigma^2 - \tau^2}{2\tau^2}$.



3.7. The Model Self-Consistency Test

We note that a more limited convergence test is possible, by reversing the procedure, and calculating the entropy of the model (which can be done directly, either analytically or by simulation). We define a self-consistency measure

$$\delta_{SC} = -H(\Psi(X)) - \overline{L_e}$$

where $H(\Psi(X))$ is the entropy of our model.

For $\Psi \to \Omega$, $\delta_{SC} \to 0$. We use this fact to construct a test

$$|-H(\Psi(X)) - \overline{L_e}| > \sqrt{\frac{\overline{Var(L_e)}}{n\epsilon}}$$

for rejecting the null hypothesis that $\Psi = \Omega$ at confidence $1 - \epsilon$.

4. Model Information

4.1. What is "Prediction"?

We defined our empirical information metric as a measure of prediction power. However, it seems worthwhile to ask again what exactly we mean by "prediction". The empirical density estimation procedure outlined above suggests that in the limit of large sample size there is always a trivial way of obtaining perfect prediction power: Copy the empirical density for X as our "likelihood model" for X, and show that it accurately predicts new observations of X. Such a procedure does not seem to qualify as "prediction"; we simply copied the observed density. In this case all the information for the "prediction" came from the observed data, and none at all from the modeling procedure itself. This suggest several conclusions:

- We desire a metric for the *intrinsic* prediction power of a model, above and beyond just copying the existing observation density. We will refer to this as I_m, the *model information*.

- Generalizing our original definition of "prediction power", we wish to maximize our prediction accuracy not only for situations that we have already observed, but also for novel situations that we have never encountered before. In other words, we adopt the conservative position that our data may be incomplete, so we cannot assume that future experience will simply mirror past experience. To maximize future prediction power, we must seek models that predict future observations more accurately than simply interpolating from past observations.

- Of course, we do not know *a priori* that such models even exist; that is a strictly empirical question. We simply generate models and measure whether they have such intrinsic prediction power, *i.e.*, $I_m > 0$.

- By definition, such a measurement can only be performed via *new observations*, e.g., a region of observation space that we have not observed before. As we will show in a moment, a region that has already been observed (thoroughly) cannot yield significant model information, because the past observations already provide a good density image for predicting future observations in this region.

- Thus, we can consider the adoption of a new model to be a *cut* on the temporal sequence of observations, partitioning them into two sets: The "old" observations (those taken before the adoption of the model), and the "new" observations (those taken after the adoption of the model).

4.2. *Defining Model Information*

The key question of model information is whether the model yields better prediction power than simple interpolation from past observations. As the interpolation reference, we simply use the empirical density calculation defined previously. Specifically, for a model Ψ we define its model information as

$$\overline{I_m(\Psi)} = \overline{L_e(\Psi|new)} + \overline{H_e(new, old)}$$

where $\overline{L_e(\Psi|new)}$ is calculated specifically using the *new* observations, and we define $\overline{H_e(new, old)} = -\overline{L_e(P_{e,old}|new)}$ as the *empirical cross entropy* of the *new* observations versus the *old* observations; $P_{e,old}$ is the empirical density estimator from the *old* observations. One example implementation (based on the previous empirical density estimator) is

$$\overline{H_e(new, old)} = -\frac{1}{n}\sum_{j=1}^{n} \log \frac{\sum_{i=1}^{n_{old}} \kappa_{X_{j,new}+\delta x/2}(X_{i,old}) - \sum_{i=1}^{n_{old}} \kappa_{X_{j,new}-\delta x/2}(X_{i,old})}{n_{old}\delta x}$$

$$\xrightarrow{LLN} -\int_{-\infty}^{\infty} \Omega(X) \log P_{e,old}(X) dX$$

where $X_{j,new}$ is the j th observation from the new observation set, $X_{i,old}$ is the i th observation from the old observations, n is the sample size of the new observations, and n_{old} is the sample size of the old observations. Many other $\overline{H_e(new, old)}$ estimation implementations are possible. It should be noted that proper normalization of the empirical density is especially important for cross-entropy calculation; however, we will not investigate such implementation details here.

Thus, $\overline{I_m}$ measures whether the model's empirical log-likelihood $\overline{L_e}$ on the *new* observations exceeds the average log-likelihood of the *new* observations computed from the *old* observation density, *i.e.*, $-\overline{H_e(new, old)}$. As for the potential information, we define a lower bound estimator for I_m with confidence level $1 - \epsilon$ based on the Law of Large Numbers:

$$I_{m,\epsilon} = \overline{I_m} - \sqrt{\frac{Var(L_e - \log P_{e,old})}{n\epsilon}}$$

- In the case $n_{old} \to 0$ we make the density function converge to the uninformative prior based on the detector range for the observable X. That is, if the range of detectable values for X is [0,10] then $P_{e,old}(X) \to 1/10$.

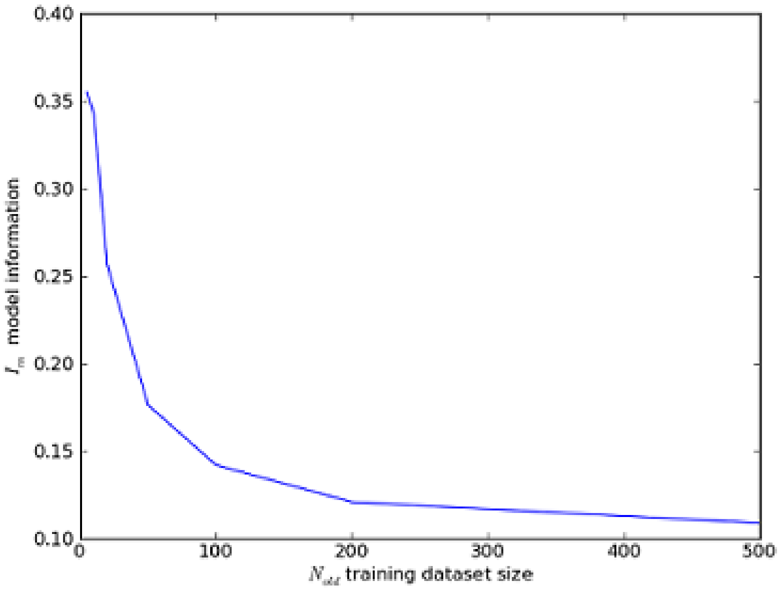
- Note that the model information can be *negative*, indicating that the model has worse prediction power than the old empirical density estimator.

4.3. *Example: The Normal Distribution*

Figure 3: Model Information of the Normal Distribution. We draw n_{old} observations from the unit normal distribution $N(0, 1)$ and compute the posterior likelihood distribution for this sample. We then

draw a new sample of 100 observations from the same distribution and use it to measure I_m for our model. The model information is initially high because the normal model predicts the shape of the distribution much more accurately than simple interpolation from the *old* observation sample.

Figure 3. Model information of the normal distribution. A model can exceed the prediction power of the empirical density computed from the training observations, because the model predicts the complete shape of the probability distribution, and how fast the tails will go to zero. Of course, as the training dataset size increases, the training data constitute a more and more accurate competing "model", and the model information decreases asymptotically. For each dataset size, a sample of that size was drawn from a unit normal distribution, and used to train a normal distribution Ψ based on the sample mean and variance. We then computed $I_m(\Psi)$ using a test sample of size 100 drawn from the unit normal. This procedure was repeated 1000 times, to obtain the average of $I_m(\Psi)$ for that training dataset size.



4.4. *Example: The Binomial Distribution*

By contrast, the binomial distribution doesn't yield significant model information, because the observable has only two possible states (*success* or *failure*) for the model to predict, and the binomial model's prediction of its probability is just equivalent to the empirical probability in the training data:

$$p(\text{success}|s_{old}, n_{old}) = \frac{s_{old} + 1}{n_{old} + 2}$$

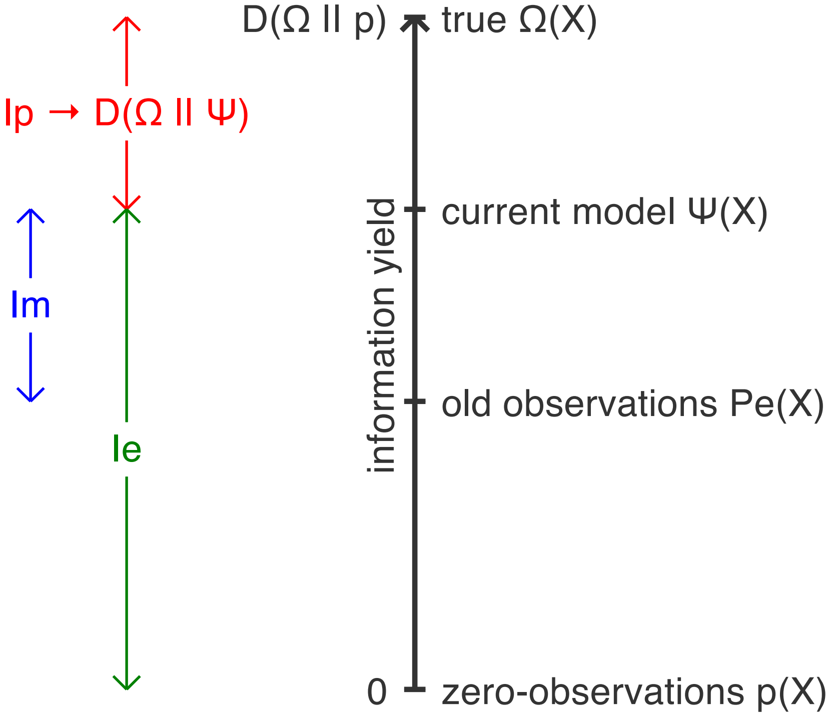
where s_{old} is the count of *successes* in the training data, and n_{old} is the size of the training data set (the +1 and +2 arise from the pseudocount principle, derived by Laplace as his "rule of succession" [19]). Fundamentally, since there is no "shape" for the model to predict (as there would be for a continuous variable, as in the case of the Normal distribution above), there is no way for the model to systematically outperform the empirical distribution.

5. Empirical Information Partition Rules

5.1. The $I_p + I_e, I_e - I_m, I_p + I_m$ Partitions

We now briefly consider the relationships between potential information, empirical information and model information, illustrated in *Figure 4: Empirical Information Partition Rules*.

Figure 4. Empirical information partition rules. This diagram illustrates the three basic partition rules: 1. **total information**: $I_p + I_e \rightarrow D(\Omega||p)$ 2. **new observations yield**: $I_p + I_m \rightarrow D(\Omega||P_{e,old})$ 3. **old observations yield**: $I_e - I_m \rightarrow D(\Omega||p) - D(\Omega||P_{e,old})$. The vertical axis represents increasing information yield, starting from zero when there are no observations, to a maximum of $D(\Omega||p)$. This axis is split by two intermediate points, the current model, $\Psi(X)$; and the old observation density $P_{e,old}(X)$. Colored intervals represent the three information metrics: I_p *(red),* I_e (green), I_m (blue).



- *All information originates as potential information.* That is, before we have a successful model for a set of observations, our prediction power is no better than random, and this manifests as positive I_p and zero I_e.
- *For a given observable X, the sum of $I_p + I_e$ is a constant (i.e., independent of the model $\Psi(X)$).* That is, for any observation sample $\vec{X}^n$,

$$\overline{I_p(\Psi)} + \overline{I_e(\Psi)} = -\overline{H_e} - \overline{L_e(\Psi)} + \overline{L_e(\Psi)} - \overline{L_e(p)} = -\overline{H_e} - \overline{L_e(p)} = \overline{I_p(p)}$$

where $p(X)$ is the uninformative distribution for X. For large sample size n

$$\overline{I_p(\Psi)} + \overline{I_e(\Psi)} = \overline{I_p(p)} \overset{LLN}{\longrightarrow} D(\Omega||p)$$

which is simply the relative entropy of the true distribution relative to the uninformative distribution $p(X)$.

- *Thus potential information is converted to empirical information by modeling.* As the model Ψ becomes a more accurate image of the observation density, I_p decreases and I_e increases by the same amount.

- *relation to mutual information*: It must be emphasized that the mutual information $I(X;\Omega)$ is defined only if we know the complete joint distribution $p(X,\Omega)$. Since we do not know this joint distribution, we would like a sampling-based estimator for $I(X;\Omega)$. We can do this by simply sampling different inference cases $\Omega_{(1)}, \Omega_{(2)}, ...\Omega_{(m)}$ (represented by different observation samples $\vec{X}^n_{(1)}, \vec{X}^n_{(2)}, ...\vec{X}^n_{(m)}$). Taking the average of $\overline{I_p(\Psi)} + \overline{I_e(\Psi)}$ over a large number m of inference cases converges:

$$\frac{1}{m}\sum\left(\overline{I_p} + \overline{I_e}\right) \overset{LLN}{\longrightarrow} -E(H(p(X|\Omega))) - E(\log p(X))$$

$$= \sum_{X,\Omega} p(X,\Omega)\log p(X|\Omega) - \sum_{X,\Omega} p(X,\Omega)\log p(X)$$

If we explicitly assume that the uninformative distribution used for computing the empirical information matches the true marginal distribution of X, then

$$= -H(X|\Omega) + H(X) = I(X;\Omega)$$

Thus, $I_p + I_e$ may be considered to be a "sampleable version of the mutual information"; that is, it can be measured for any individual inference case, and its average over multiple inference problems will converge to the mutual information of the observable *vs.* hidden variables.

- *For a given observable X, the sum of $I_e - I_m$ is a constant. (i.e., independent of the model $\Psi(X)$).* Assuming both I_e, I_m are calculated on the same test data,

$$\overline{I_e(\Psi)} - \overline{I_m(\Psi)} = \overline{L_e(\Psi)} - \overline{L_e(p)} - \overline{L_e(\Psi)} + \overline{L_e(P_{e,old})} = -\overline{L_e(p)} + \overline{L_e(P_{e,old})}$$

where $P_{e,old}(X)$ is the distribution of X computed from past observations (as described above). So for $n \to \infty$

$$\overline{I_e(\Psi)} - \overline{I_m(\Psi)} \overset{LLN}{\longrightarrow} D(\Omega||p) - D(\Omega||P_{e,old})$$

Thus $I_e - I_m$ measures the amount of information supplied by the past observations (in the form of $P_{e,old}(X)$).

- Moreover, in the asymptotic limit, $\overline{I_e} - \overline{I_m} \geq 0$ since for $n_{old} \to 0$ we guarantee that $P_{e,old}(X) \to p(X)$ and for $n_{old} \to \infty$ we have $P_{e,old}(X) \overset{LLN}{\longrightarrow} \Omega(X)$.

- Thus, I_m partitions I_e into the part that is simply provided by the training observations themselves, versus the part that actually constitutes "value added" predictive power of the model itself.

- *For a given observable X, the sum of $I_p + I_m$ is a constant (i.e., independent of the model $\Psi(X)$).* Specifically, assuming both I_p, I_m are calculated on the same test data,

$$\overline{I_p(\Psi)} + \overline{I_m(\Psi)} = -\overline{H_e} - \overline{L_e(\Psi)} + \overline{L_e(\Psi)} - \overline{L_e(P_{e,old})} = \overline{I_p(P_{e,old})}$$

$$\xrightarrow{LLN} D(\Omega||P_{e,old})$$

  which simply measures the amount of information available to be learned about the true distribution of X above and beyond that already provided by past observations (in the form of $P_{e,old}(X)$).

- *Relation of I_m to relative entropy*: Note that since $I_p \xrightarrow{LLN} D(\Omega||\Psi)$, this also implies that $I_m \xrightarrow{LLN} D(\Omega||P_{e,old}) - D(\Omega||\Psi)$. This simply restates the principle that the model information represents the increase in model prediction power relative to the empirical density of the past observations.

5.2. *Asymptotic Conversion of Potential and Model Information to Empirical Information*

Consider the following asymptotic modeling protocol: For a large sample size $n_{old} \to \infty$ we simply adopt the empirical density $P_{e,old}$ as our model Ψ. We then measure I_e, I_p, I_m on a set of *new* observations.

As $n_{old} \to \infty$, $P_{e,old}(X)$ converges to the true density $\Omega(X)$, so $H_e(new, old) \xrightarrow{LLN} H(\Omega(X))$ and

$$\overline{I_m} \xrightarrow{LLN} -H(\Omega(X)) - D(\Omega||\Psi) + H(\Omega(X)) = -D(\Omega||\Psi) \leq 0$$

Since the relative entropy is non-negative, the maximum attainable value of the model information drops asymptotically to zero. Moreover, as $\Psi(X) = P_{e,old}(X)$ also converges to the true density $\Omega(X)$, $\overline{I_p} \xrightarrow{LLN} D(\Omega||\Omega) = 0$. Since both the model and potential information vanish, by the $I_p + I_e$ and $I_e - I_m$ partition rules, all information is converted exclusively to empirical information.

This scenario illustrates a simple point about the distinct meanings of empirical information *vs.* model information. The overriding goal of model selection is maximizing empirical information (likelihood). However, this scenario shows that maximizing the empirical information is in a sense trivial if one can collect a large enough observation sample. By contrast, there is no trivial way to produce positive model information; note that the very procedure that automatically maximizes I_e also ensures that $I_m \leq 0$.

This suggests several changes in how we think about the value of modeling. In model selection, the value of a model is often thought of in terms of data compression; that is, that the best model encodes the underlying pattern of the data in the most efficient manner possible. Metrics such as the AIC and BIC seek to enforce this principle by adding "correction terms" that penalize the number of model parameters. However, to be truly valuable for prediction, a model should meet this data compression criterion not only retrospectively (*i.e.*, it can yield a more efficient encoding of the past observations) but also prospectively (*i.e.*, it can predict future observations more accurately than simply interpolating from the past observations). Whereas the total empirical information metric fails to draw this distinction, the model information explicitly measures it. That is, it partitions the total I_e into a "trivial" part that represents the prediction power implicit in the observation dataset itself, and a non-trivial part that represents true "predictions" coming from the model.

6. Conclusion

We wish to suggest that these empirical information metrics represent a useful extension of existing statistical inference metrics, because they provide "sampleable" measures of key information theory metrics (such as mutual information and relative entropy), with explicit Law of Large Numbers convergence guarantees. That is, each empirical information metric can be measured via sampling on an individual inference problem (unlike the conventional definition of mutual information); Yet its average value over multiple inference problems will converge to the true, hidden value of its associated metric from information theory (such as the mutual information). On such a foundation, one can begin to recast statistical and scientific inference problems in terms of the very useful and general tools of information theory. For example, the "inference halting problem", which imposes a variety of problems and limitations in Bayesian inference, can be easily resolved by the potential information metric, which directly measures the distance of the current model from the true distribution in standard information theoretic terms. Similarly, the model information metric measures the "value-added" prediction power of a model relative to its training data.

Acknowledgements

The author wishes to thank Marc Harper, Esfan Haghverdi, John Baez, Qing Zhou, Alex Alekseyenko, and Cosma Shalizi for helpful discussions on this work. This research was supported by the Office of Science (BER), U. S. Department of Energy, Cooperative Agreement No. DE-FC02-02ER63421.

References

1. Shannon, C. A Mathematical Theory of Communication. *Bell System Tech. J.* **1948**, *27*, 379–423.
2. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley: New York, NY, USA, 1991.
3. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **1974**, *AC-19*, 716–23.
4. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
5. de Finetti, B. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **1937**, *7*, 168.
6. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
7. Efron, B. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **1981**, *68*, 589–599.
8. Breiman, L. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Am. Stats. Assoc.* **1992**, *87*, 738–754.
9. Geisser, S. *Predictive Inference*; Chapman and Hall: New York, NY, USA, 1993.
10. McQuarrie, A.; Tsai, C.L. *Regression and Time Series Model Selection*; World Scientific: Singapore, 1998.

11. Shalizi, C.R. Dynamics of Bayesian Updating with Dependent Data and Misspecified Models. *Electron. J. Statist.* **2009**, *3*, 1039–1074.
12. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis, 2nd ed.*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2003.
13. Bonnlander, B.; Weigend, A. Selecting input variables using mutual information and nonparametric density estimation. Proceedings of the 1994 International Symposium on Artificial Neural Networks (ISANN 94); Taiwan, 1994; pp. 42–50.
14. Kraskov, A.; Stogbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138.
15. Kullback, S.; Leibler, R. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
16. Sawa, T. Information Criteria for Discriminating among Alternative Regression Models. *Econometrica* **1978**, *46*, 1273–1291.
17. Vuong, Q. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **1989**, *57*, 307–333.
18. Paninski, L. Asymptotic theory of information-theoretic experimental design. *Neural Computat.* **2005**, *17*, 1480–1507.
19. Laplace, P.S. *Essai philosophique sur les probabilités*; Courcier: Paris, France, 1814.

© 2011 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/.)