Information Theory and Data Analysis

These notes include portions of "Entropy and Information" notes from 2003-2004

The odd title is emphasize at the outset that the plan here is not to suggest that information theory represents an organizing principle for the brain – but just to present some ideas from information theory that are useful in data analysis. The main points are that "mutual information" is a principled, nonparametric measure of whether there is a statistical dependence between two variables, and that "maximum entropy" distributions are useful, familiar objects that formalize ignorance about what we don't measure.

Entropy

We begin with developing a way to measure the richness of a distribution, "entropy." More precisely, we want to quantify how difficult it is to specify a value that is chosen from the distribution – the richer (more varied) the distribution, the more difficult it is to specify a value.

What about the variance, $M_2 = \left\langle \left(x_i - \langle x \rangle \right)^2 \right\rangle = \sum p_i (x_i - \langle x \rangle)^2$ (where $\langle x \rangle = \sum p_i x_i$)?
The problem with using variance is that it is parametric. That is, it depends on the *values* that are associated with each probability (i.e., the values x_i, each of which are drawn with probability p_i) that are drawn , not just the probabilities themselves. If we change the units of the values, then we change the variance – but we don't change the difficulty of specifying a value. Same problem if we use a higher moment, such as $M_4 = \left\langle \left(x_i - \langle x \rangle \right)^4 \right\rangle$. We could get rid of the dimensions by considering instead the kurtosis, $\kappa = \frac{M_4}{\left(M_2 \right)^2} - 3$ (the "3" is because $M_4 / {M_2}^2 = 3$ for a Gaussian, more on the privileged role of Gaussians below), but this still depends on the values x_i. What about functions *f* that operate directly on the set of probabilities themselves? There are still lots of possibilities – we could use the variance (or higher moments) of the probabilities $p_1, \ldots, p_m$ (with $\sum p_i = 1$, and each $p_i \geq 0$) to measure how evenly distributed they are, etc. So there are lots of possibilities for *f*.

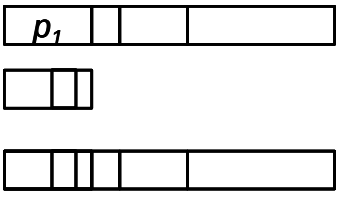
Axiomatic approach

But if we agree on some properties for *f*, we can essentially narrow down the possibilities to just one. The axioms are intended to codify the notion that *f* measures how many yes/no questions must be asked, on average, to determine the value drawn. (There are other axiomatic approaches; this is one of the simplest.)

The first axiom is that the entropy of independent distributions must add. Formally: Given a distribution *X* that assigns the probabilities $p_1, \ldots, p_m$ to the symbols $\{1, \ldots, m\}$, and a distribution

Y that assigns the probabilities $q_1,\ldots,q_n$ to the symbols $\{1,...,n\}$, we can create a new ("direct product") distribution $X \times Y$ that assigns the probability $p_j q_k$ to a symbol (j,k), with j in $\{1,...,m\}$ and k in $\{1,...,n\}$. We require that $f(p_1q_1, p_1q_2,\ldots, p_jq_k) = f(p_1,\ldots,p_m) + f(q_1,\ldots,q_n)$. Put another way, the specifying a member of $X \times Y$ is the same as specifying a member of X, and then a member of Y.

The second axiom is that if we start with one distribution, and then refine it into a second one, we can extend the calculation in the intuitive way. This is known as the "chain rule", and it is easier to illustrate than to formalize. Here, X is the initial distribution (on m symbols), and the first symbol, which occurs with probability p_1, can be refined into n distinguishable symbols, as each with probability q_k. The q_k's are conditional probabilities, i.e., they are the probabilities of refinement of X's first symbol into each of the n symbols of Y, given that this first symbol occurs. So $\sum q_i = 1$, and the probability of each of these refined symbols in X is $p_1 q_k$



$f(X)$

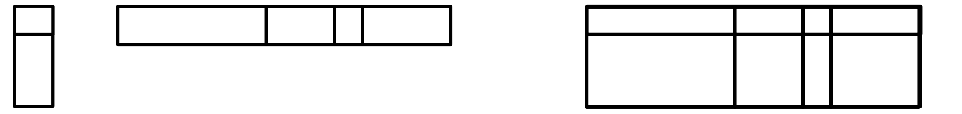
$f(Y)$

$f(Z) = f(X) + p_1 f(Y)$

Given this setup, the chain rule is formalized by

$$f(p_1q_1, p_1q_2,\ldots, p_1q_n, p_2, p_3,\ldots, p_m) = f(p_1,\ldots,p_m) + p_1 f(q_1,\ldots,q_n).$$

Interestingly, the chain rule axiom implies the additivity axiom.



$f(X)$ $f(Y)$ $f(X \times Y) = f(X) + f(Y)$

The way to see this to consider the distribution X as the "unrefined" distribution, and then, successively refine each symbol by the distribution Y.
Refining the first symbol of X, which occurs with probability p_1:

$$f(p_1q_1, p_1q_2,\ldots, p_1q_n, p_2, p_3,\ldots, p_m) = f(p_1,\ldots,p_m) + p_1 f(q_1,\ldots,q_n)$$

Refining the second symbol of X, which occurs with probability p_2:

$$f(p_1q_1, p_1q_2,\ldots, p_1q_n, p_2q_1, p_2q_2,\ldots, p_2q_n, p_3,\ldots, p_m)$$
$$= f(p_1q_1, p_1q_2,\ldots, p_1q_n, p_2, p_3,\ldots, p_m) + p_2 f(q_1,\ldots,q_n),$$
$$= f(p_1, p_2, p_3,\ldots p_m) + (p_1 + p_2) f(q_1,\ldots,q_n)$$

etc. Once all the symbols have been refined, we recover the additivity rule:

$$f(p_1q_1,\ldots,p_mq_n) = f(p_1,p_2,p_3,\ldots,p_m) + (p_1+p_2+\ldots+p_m)\,f(q_1,\ldots,q_n)$$
$$= f(p_1,\ldots,p_m) + f(q_1,\ldots,q_n)$$

The axiom(s) determine *f* uniquely, up to a multiplicative constant. (Actually, we need to add one more axiom, that *f* depends continuously on its arguments. This is meant in a strong way -- if we consider the argument of *f* as a vector of probabilities, then the value assigned by *f* is a continuous function of that vector.)

First, considering a "trivial" distribution consisting of only one symbol, we must have $f(1)=0$. (This is because we can take *Y* to be this distribution and apply the chain rule: after this trivial "refinement", $Z = X$; since also, $Z = X + p_1 f(1)$, it follows that $f(1)=0$.

Next, we consider the next-least-trivial distribution, an equal bipartite distribution *B*, $p_1 = p_2 = 1/2$, and we arbitrarily assign $f(1/2,1/2) = a$ for some constant *a*. Applying the additivity rule N times (i.e., creating an *N*-fold product $B\times\ldots\times B$) yields $f(2^{-N},\ldots,2^{-N}) = Na$. Put another way, $f(1/M,\ldots,1/M) = -a\log_2(M)$, provided that *M* is a power of 2.

We could now take the same approach, beginning with a distribution on *K* items, for which $p_1 = \ldots = p_K = 1/K$. We assign $f(1/K,\ldots,1/K) = a_K$. Iterating the additivity rule yields $f(K^{-N},\ldots,K^{-N}) = Na_K$, or, $f(1/M',\ldots,1/M') = -a_K\log_K(M') = -a_K\dfrac{\log_2(M')}{\log_2(K)}$, provided that M' is a power of *K*.

We'd like to relate a_K to *a*. If there were an integer M'' that was both a power of 2 and a power of *K*, we could write $-a_K\dfrac{\log_2(M'')}{\log_2(K)} = -a\log_2(M'')$, from which it would follow that $a_K = a\log_2(K)$. But typically (unless *K* itself is a power of 2)l there's no such integer M''. Nevertheless, the conclusion that $a_K = a\log_2(K)$ must hold. The reason is the following. We can find an integer power of 2 and an integer power of *K* that are *arbitrarily close* to each other (in a ratio sense). This allows us to conclude (by the assumption that *f* is continuous) that the ratio between a_K and $a\log_2(K)$ is arbitrarily close to 1. So, it follows that for any *K*, $f(1/K,\ldots,1/K) = a\log_2 K$.

Finally, we need to consider the case in which the arguments of *f* are unequal. We choose a large denominator *D*, so that the probabilities $p_1,\ldots p_m$ are as close as desired to integer multiples of $1/D$, i.e., $p_i \approx k_i/D$. We now compute $f(p_1,\ldots p_m)$ by using the "chain rule" in reverse. That is, we calculate $f(1/D,\ldots,1/D) = a\log_2 D$, and then group together the first k_1 arguments to make (approximately) p_1, the next k_2 arguments to make (approximately) p_2, etc. The *i*th refinement has k_i equal arguments, so it contributes $a\log_2 k_i$. It follows that

$$f(p_1,\ldots p_m)+\sum_i p_i a\log_2 k_i \approx a\log_2 D = a\sum_i p_i \log_2 D\text{, or,}$$

$$f(p_1,\ldots p_m)\approx a\sum_i p_i(\log_2 D-\log_2 k_i)\approx a\sum_i p_i \log_2 \frac{D}{k_i} = -a\sum_i p_i \log_2 p_i\ .$$

Definition of entropy

This shows that, up to a multiplicative constant *a*, there is a unique definition of *f* that satisfies these very simple axioms, which we call the "entropy." Typically, we take $a=1$, so that $f(1/2,1/2)=1$, yielding the entropy in bits -- the number of binary yes/no questions required to specify, on average, a sample from the distribution. Thus, the entropy of a discrete distribution *P* is defined by

$$H(P)=-\sum_i p_i \log_2 p_i\ . \qquad (1)$$

Note that the order of the symbols does not matter – they are simply abstract labels.

We also note that we need to show that the quantity defined by eq. (1) satisfies the axioms for any distribution *P* , and not just for the ones needed to determine the definition of entropy. This is straightforward (and one only needs to show that eq. (1) satisties the chain rule).

While the entropy is not a moment of the distribution of probabilities, it is "almost" a moment. This viewpoint is expressed by the following restatement of the above definition:

$H(P)=-\frac{1}{\ln 2}\frac{d}{d\alpha}\sum_i (p_i)^\alpha\Big|_{\alpha=1}$. (This is equivalent to eq. (1) because $\frac{d}{d\alpha}p^\alpha=\frac{d}{d\alpha}e^{\alpha\ln p}=e^{\alpha\ln p}(\ln p)=p^\alpha\ln p$.) Via L'Hopital's rule, this has another equivalent form:

$$H(P)=\frac{1}{\ln 2}\lim_{\alpha\to 1}\frac{\sum_i (p_i)^\alpha-1}{1-\alpha}\ .$$

The numerical factor $1/\ln 2$ in the above would disappear had we chosen $a=\ln 2=\log_e 2$, yielding the entropy in "nats".

Mixing and convexity

Here we'll demonstrate a basic properties of entropy – that ultimately enable us to turn it into "mutual information," a nonparameteric measure of associaton. These properties also justify

using the term “information” for these measures, and also are the basic reason that entropy is tricky to estimate from data.)

Say we mix two distributions, *P* and *Q*, taking a proportion *z* from *P* and $1-z$ from *Q*. The new distribution $R_z = (1-z)P + zQ$ corresponds to drawing from *P* on a fraction *z* of the trials, and drawing from *Q* on the remaining $1-z$ of the trials.

We’ll show that $H(R_z)$ is a concave-down function of *z*. That is, $H(R_z)$ lies at or above the line segment that runs from $H(P)$ at $z=0$ to $H(Q)$ at $z=1$.

This is straightforward calculus:

$R_z = (1-z)P + zQ$, so

$$H(R_z) = -\frac{1}{\ln 2}\sum_i \big((1-z)p_i + zq_i\big)\ln\big((1-z)p_i + zq_i\big),$$

so

$$\frac{d}{dz}H(R_z) = -\frac{1}{\ln 2}\sum_i (q_i - p_i)\Big(1+\ln\big((1-z)p_i + zq_i\big)\Big) = -\frac{1}{\ln 2}\sum_i (q_i - p_i)\Big(\ln\big((1-z)p_i + zq_i\big)\Big),$$

(last equality since $\sum p_i = \sum q_i = 1$), and

$\frac{d^2}{dz^2}H(R_z) = -\frac{1}{\ln 2}\sum_i \frac{(q_i - p_i)^2}{(1-z)p_i + zq_i}$. Since this quantity must be negative, $H(R_z)$ is a concave-down function of *z*.

For $z = 1/2$, this means that “the entropy of the average is greater than the average of the entropies.”

The downward bias of entropy estimates

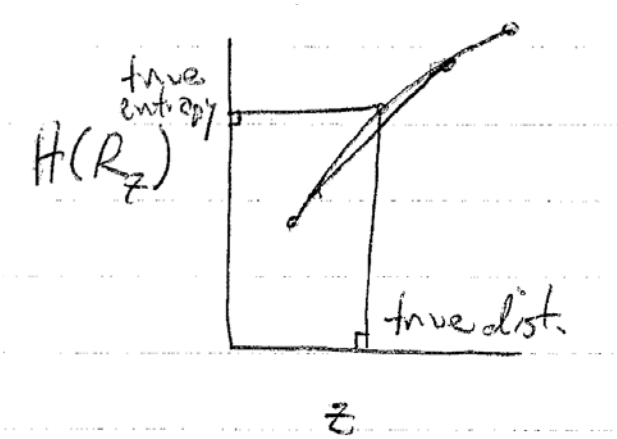
The convexity property has an important consequence for the practical issue of how to estimate entropy from data. The obvious way to proceed is to estimate the probabilities of each symbol from the empirical dataset (e.g., if the symbol *i* occurs n_i times in N observations, assume that $p_i = n_i / N$), and then to plug these probabilities into eq. (1). The convexity property implies that this estimator will be downwardly biased.

The reason is the following. The true values of the probabilities are their values in an ideal, infinite ensemble, of which every real-world experiment is a finite sample. The ideal ensemble is, exactly, the mixture of all of the finite samples. So, the entropy of the infinite ensemble must be greater than the mixture (i.e., average) of the estimates from finite samples.

What is not so obvious is that the size of the error is (asymptotically for large *N*) given by $(K-1)/(N\ln 2)$, where K is the number of different letters. Since this is independent of the distribution, it would seem that one can simply correct the naïve estimate by this amount. This strategy is predicated on knowing the number of different letters, rather than estimating it. And

there's more fine print: if the number of samples of each symbols is not large, then using the debiaser, which is only *asymptotically correct*, may result in an estimate that is worse than the naïve one.

There are more sophisticated ways to deal with this problem – estimators that improve on the bias, at the expense of making assumptions about the distribution. The more specific the assumptions, the better the estimator (provided that the assumptions are correct.)



Continuous distributions

The notion of entropy can be extended to continuous distributions, e.g., specified by $p(x)$, where $\int_X p(x)dx = 1$, and $p(x)\Delta x$ is the probability that a draw from X yields a value within some interval (or volume) Δx centered at x.

The natural extension of eq. (1) to this scenario is

$$H = -\int_X p(x)\log_2(p(x))dx\,. \qquad (2)$$

There is an important piece of fine print. If you approximate a continuous distribution by a finely spaced discrete distribution and find the entropy of each (using eq. (2) for the continuous distribution, and eq. (1) for the discrete approximation, with $p_i = p(x_i)\Delta x$), you will not get the same answer:

$$H_{continuous} = -\int_X p(x)\log_2(p(x))dx \approx -\sum_i p_i \log_2(p_i/\Delta x) = -\sum_i p_i \log_2(p) + \log_2(\Delta x)$$
$$= H_{discrete} + \log_2(\Delta x)$$

Thus, in the continuous case, there is an arbitrary offset inherent in the definition of eq. (2), effectively equal to the log of the volume element. So in the continuous case, absolute entropies are not very interpretable – but differences in entropies (measured with the same units) always

make sense. As a reminder of the distinction between the continuous case and the discrete case, the quantity defined by eq. (2) is known as the "differential entropy."

Mutual information

Now let's say we have two sets of symbols, *X* and *Y*, that may have some statistical dependence. For example, they may represent the input and output of a data channel; if the data channel is a good one, there will be strong dependence. We can consider quantifying the degree of dependency between them in three ways:

Q_Y : To what extent does knowing the symbol *y* drawn from Y reduce the number of yes/no questions required to determine the symbol *x* drawn from *X*?

Q_X : To what extent does knowing the symbol *x* drawn from *X* reduce the number of yes/no questions required to determine the symbol *y* drawn from *Y*?

Q_{XY} : To what extent does one need fewer yes/no questions to specify a typical (x, y)-pair, than to separately specify *x* and *y*?

It turns out that all of these yield the same quantity, which is known as the mutual information between *X* and *Y*, denoted $I(X,Y)$. This quantity is necessarily 0 or positive, and it is 0 only when *X* and *Y* are statistically independent.

We now demonstrate this.

We write $p(x)$ as the probability that a symbol *x* is drawn from *X*; $p(x, y)$ as the probability that the pair (x, y) is drawn, and $p(x \mid y)$ as the probability that *x* is drawn from *X*, given that *y* is drawn from *Y* (and similarly $p(y)$ and $p(y \mid x)$). So, for example, $p(x) = \sum_y p(x, y)$, and $p(x \mid y) = p(x, y) / p(y)$. We also write $X \mid y$ as the conditional distribution of *X*, given a particular draw *y*, and similarly for $Y \mid x$. With these notations,

$$Q_X = H(Y) - \sum_x p(x) H(Y \mid x),$$

$$Q_Y = H(X) - \sum_y p(y) H(X \mid y),$$

and

$$Q_{XY} = H(X) + H(Y) - H(X,Y).$$

It suffices to show that $Q_X = Q_{XY}$, since Q_{XY} is symmetric in *X* and *Y*. Using the above, $Q_X = Q_{XY}$ is equivalent to $-\sum_x p(x) H(Y \mid x) = H(X) - H(X,Y)$, or, $H(X,Y) = H(X) + \sum_x p(x) H(Y \mid x)$. The latter follows from the chain rule: we view a symbol pair (x, y) as the symbol *x*, refined by learning that this is associated with *y*.

For reference, we rewrite Q_Y, Q_X, and Q_{XY} (all of which are equal to $I(X,Y)$) in terms of the probabilities themselves:

$$Q_Y = -\sum_x p(x)\log_2 p(x) + \sum_y p(y)\sum_y p(x\,|\,y)\log_2 p(x\,|\,y)$$
$$= -\sum_x p(x)\log_2 p(x) + \sum_{x,y} p(x,y)\log_2 p(x\,|\,y) \quad ,$$

$$Q_X = -\sum_y p(y)\log_2 p(y) + \sum_x p(x)\sum_y p(y\,|\,x)\log_2 p(y\,|\,x)$$
$$= -\sum_y p(y)\log_2 p(y) + \sum_{x,y} p(x,y)\log_2 p(y\,|\,x) \quad ,$$

$$Q_{XY} = -\sum_x p(x)\log_2 p(x) - \sum_y p(y)\log_2 p(y) + \sum_{x,y} p(x,y)\log_2 p(x,y) \, .$$

Note that, as with entropy, the order of the symbols x_i and y_i does not matter – they are simply abstract labels.

The "fine print" that arises when entropy is measured for a continuous distribution does not apply here – since the "$\log_2(\Delta x)$" terms cancel (e.g., in the definition Q_Y). For Q_{XY}, this cancellation occurs because $H(X,Y)$ has an additive offset of $\log_2(\Delta x \Delta y) = \log_2(\Delta x) + \log_2(\Delta y)$, and each of the individual log-terms on the right cancel with the additive offsets of $H(X)$ and $H(Y)$.

Nonparametric measure of statistical dependence

An important property of mutual information is that $I(X,Y) = 0$ when *X* and *Y* are independent, and $I(X,Y) > 0$ otherwise (i.e., they are statistically dependent). Note that here we're talking about something much more general than "correlation" in the sense of $\left\langle \left(x - \langle x \rangle\right)\left(y - \langle y \rangle\right) \right\rangle$: I measures something that is entirely nonparametric, and *x* and *y* are simply abstract symbols.

First, note that if *X* and *Y* are independent, then $p(x,y) = p(x)p(y)$ and the additivity property immediately implies that $Q_{XY} = 0$.

Conversely, if *X* and *Y* are not independent, we will use the mixing property to show that $Q_X > 0$. If *X* and *Y* are not independent, then there is at least one pair of symbols x_1 and x_2 for which $p(y\,|\,x_1) \neq p(y\,|\,x_2)$. Now consider replacing $p(y\,|\,x_1)$ and $p(y\,|\,x_2)$ by their weighted sums, $\dfrac{p(y\,|\,x_1)p(x_1) + p(y\,|\,x_2)p(x_2)}{p(x_1) + p(x_2)}$. This is the same as scrambling the symbols symbols x_1 and x_2; each still occurs the same number of times but their selective co-occurrences with *Y* are

mixed. Doing this does not change the distributions *X* or *Y*, but it does change the contribution of the terms involving x_1 and x_2 in Q_X. In particular, it mixes $p(Y \mid x_1)$ and $p(Y \mid x_2)$, with mixing parameter $z = \frac{p(x_2)}{p(x_1) + p(x_2)}$. Thus, the contribution $p(x_1)H(Y \mid x_1) + p(x_2)H(Y \mid x_2)$ increases, and Q_X decreases. Such mixings can be continued as long as there are any symbols in *X* for which the conditional probabilities differ; each mixing further decreases Q_X. Finally, when all conditional probabilities have been equated by mixing, *X* and *Y* are independent, and, as above, $Q_X = Q_{XY} = 0$.

Data processing inequality

The same kind of reasoning leads to another important property of mutual information, the "data processing inequality." Informally, it states that if *Z* is derived from *Y*, then $I(X,Z) \leq I(X,Y)$.

We imagine that an observer is trying to reduce the uncertainty (the number of yes/no questions) that must be asked to determine a draw from *X*. But instead of using $p(X \mid y)$, corresponding to the symbol *y* that emerges from the communication channel, the observer "processes" y in some way (independent of *X*), to try to do better. This "processing" could consist of some combination of mixing in the probability $p(X \mid y')$ associated with some other symbol in *Y*, or relabeling the symbols. Relabeling does not change information, and the mixing operation (as above) can only decrease it.

Upward bias of mutual information estimates

Just as the "plugin" estimate of entropy is downwardly biased, the "plugin" estimate of information is upwardly biased. This is readily seen from the Q_{XY}-formulation for mutual information: the upward bias of the plugin estimator, asymptotically, is $\frac{K_{XY} - K_X - K_Y + 1}{N \ln 2}$, where K_X is the number of different kinds of *X*-symbols, K_Y is the number of different kinds of *Y*-symbols, and K_{XY} is the number of different (x, y)-pairings that actually occur. Debiasing strategies for entropy estimates can all be applied to information estimates, and there are other possibilities too, based on expressions for mutual information that are not explicit differences in entropies.

For linear systems with additive, Gaussian noise, there are classical results (due to Shannon) that relate the mutual information to the spectrum of the noise and the shape of the transfer function, and to the coherence between input and output.

Constrained maximum-entropy distributions

Many familiar probability distributions are "constrained maximum entropy" distributions – distributions that have the greatest possible entropy subject to some constraint, such as their mean, their variance, etc. Such distributions can be thought of as representing principled hypotheses about a statistical ensemble, based on some parametric measurements (e.g., the sample mean and variance) – and "constrained maximum entropy" formalizes the idea of finding the distribution that is as random as possible, given the measured constraints.

Constrained maximum-entropy distributions also constitute an approach to dimensional reduction. If a limited set of measurements, plus maximum-entropy, suffice to predict the entire distribution, then the limited set of measurements is a concise description of the entire distribuiton. See for example Shlens et al., J. Neurosci., 2006, and Nirenberg and Victor, Current Opinions in Neurobiology 2007, for a review.

Not surprisingly, this formulation leads to many familiar distributions: Gaussian distributions, Poisson distributions, Gaussian noise ensembles, and many others (including von Mises distributions for angular data, and Markov processes).

In typical useful cases, the constraints are linear in the probabilities. For example, given a set of numerically-valued symbols x_i that are distributed according to $p_i = p(x_i)$, the mean of a distribution is $M_1 = \sum x_i p(x_i)$. This constraint is linear in each $p(x_i)$, as each $p(x_i)$ is weighted by x_i. Similarly, any moment is a linear constraint; the *k*th moment M_k is given by $M_k = \sum (x_i)^k \, p(x_i)$; each $p(x_i)$ is weighted by $(x_i)^k$. If the mean M_1 is known or constrained, then the variance is also a linear constraint, since the variance of a distribution is $V = \sum (x_i - M_1)^2 \, p(x_i)$ -- here, each $p(x_i)$ is weighted by $(x_i - M_1)^2$.

For multivariate distributions, many useful constraints are also linear in the probabilities. The constraint that a joint distribution $p(x, y)$ has marginals $p(x') = \sum_y p(x', y)$ is a *set* of linear constraints: for the constraint corresponding to each x', the coefficient of $p(x, y)$ is 1 if $x = x'$ and 0 otherwise. Viewing a time series as a sample of a multivariate distribution, the autocovariance, and the power spectrum, are also constraints that are linear in the probabilities (for specified mean).

Note that the mixing property guarantees that for linear constraints, the constrained maximum-entropy distribution is unique. For if there had been two local maxima, then the entropy along the path that joins them ($R_z = (1-z)P + zQ$) could not be convex.

Lagrange Multipliers

The method of Lagrange Multipliers makes it easy to see how linear constraints determine maximum-entropy distributions. Since this is a generally useful piece of machinery, we take a brief detour. Lagrange Multipliers also play a central role in PCA.

Say you have a function of *m* variables, e.g., $H(p_1, p_2, \ldots, p_{m-1}, p_m)$ that you want to maximize, but that variables p_i must satisfy a constraint, e.g., $C(p_1, p_2, \ldots p_m) = b$. The straightforward way to solve this is to use the constraint equation to write one of the p_i's in terms of the others, e.g., $p_m = f(p_1, \ldots, p_{m-1})$, and then to consider $H(p_1, p_2, \ldots, p_{m-1}, f(p_1, p_2, \ldots, p_{m-1}))$ as a function of $m-1$ variables, and then to solve, simultaneously, the system

$$\frac{\partial}{\partial p_j} H(p_1, p_2, \ldots, p_{m-1}, f(p_1, p_2, \ldots, p_{m-1})) = 0, \tag{3}$$

for the $m-1$ independent variables $p_1, \ldots, p_{m-1}$.
Lagrange multipliers provides an alternative solution, that (a) often is easier, especially if *f* is difficult to write explicitly, (b) does not "single out" one of the variables p_i (and hence preserves symmetry), and (c) often provides additional insight.

The Lagrange Multiplier recipe is to replace the above constrained minimization problem by the unconstrained minimization of $L(p_1, p_2, \ldots, p_m) = H(p_1, p_2, \ldots, p_m) + \lambda C(p_1, p_2, \ldots, p_m)$, and to adjust λ so that the extremum for *L* satisfies the constraints. λ is known as a "Lagrange Multiplier." That is, the system of $m-1$ equations (3) is replaced by the system

$$\begin{cases} \dfrac{\partial}{\partial p_j}\big(H(p_1, p_2, \ldots, p_m) - \lambda C(p_1, p_2, \ldots, p_m)\big) = 0 \\ \\ C(p_1, p_2, \ldots, p_m) = b \end{cases}. \tag{4}$$

In practice, the first set of $m-1$ equations yields the unknowns p_i as functions of λ, and constraint equation becomes the "hard" part of the problem, which determines λ implicitly, via $C\big(p_1(\lambda), p_2(\lambda), \ldots, p_m(\lambda)\big) = b$.

Why does this work? We can think of $H(p_1, p_2, \ldots, p_m)$, without constraints, as defining a surface over the domain of the *p*'s. At any point on the surface, the slope of the tangent plane in any coordinate axis p_i is given by $\partial H / \partial p_i$; it is convenient to think of these slopes as a vector, the gradient of *H*: $\nabla H = (\partial H / \partial p_1, \ldots, \partial H / \partial p_1)$. (Actually, ∇H is a member of the dual space.) For a point $\vec{q} = (q_1, \ldots q_m)$ to be an extremum of the *unconstrained* problem, it would be necessary that any small movement around $\vec{q}$ does not change the value of *H*. Familiarly, this is equivalent to $\nabla H = 0$.

But for the *constrained* problem, it is OK if a small movement around $\vec{q}$ changes the value of *H*, provided that it also changes the value of the constraint! That is, for the constrained problem, the only movements around $\vec{q}$ that are allowed are the ones that keep the constraint unchanged. The small movements $\Delta\vec{p}$ that keep the constraint *C* unchanged are the ones that are orthogonal to the constraint's gradient, i.e., the directions for which $\Delta\vec{p} \cdot \nabla C = 0$. So for the constrained problem, a point $\vec{q}$ is an extremum if, at this point, any direction that is not orthogonal to ∇H (the directions in which *H* will change) are also not orthogonal to ∇C (the directions in which

the constraint will change). If the directions that are not orthogonal to ∇H match those that are not orthogonal to ∇C, then the directions that are orthogonal must match as well. This is the first condition of the set (4). The second condition of the set (4) is simply that the constraints are satisfied.

Lagrange Multipliers also work when there are multiple constraints, $C_r(p_1, p_2, \ldots p_m) = b_r$. (The above argument extends; each constraint gets its own multiplier.) The conditions for an extremum are:

$$\begin{cases} \dfrac{\partial}{\partial p_j}\left(H(p_1, p_2, \ldots, p_m) - \sum_r \lambda_r C_r(p_1, p_2, \ldots, p_m) \right) = 0 \\ C_r(p_1, p_2, \ldots, p_m) = b_r \end{cases} . \qquad (5)$$

Finding a maximum-entropy distribution subject to linear constraints

With Lagrange Multipliers, it is straightforward to write a formal solution to the problem of finding the maximum entropy distribution subject to linear constraints. We simply use the system (5), with C_r representing the constraints, and $H(p_1, p_2, \ldots, p_m) = -\sum p_i \ln p_i$. (It doesn't matter which base we use for the logs, since this just changes the entropy by a scale factor.) There's always an implicit constraint, that the *p*'s are a probability distribution – and we call this the constraint C_0, with $C_0(p_1, \ldots p_m) = \sum p_i$ and $b_0 = 1$.

Since we are assuming that the constraints are linear, each can be put in the form $C_r(p_1, \ldots p_m) = \sum_i c_{ri} p_i$. Using this and the definition of entropy for *H*, the first part of the system (5) becomes $\dfrac{\partial}{\partial p_j}\left(-\sum_i p_i \ln p_i - \sum_r \lambda_r \sum_i c_{ri} p_i \right) = 0$, or, $-1 - \ln p_j - \sum_r \lambda_r c_{rj} = 0$. Recognizing the special rule of $r = 0$ (i.e., that $c_{0j} = 1$), this becomes $-1 - \ln p_j - \lambda_0 - \sum_{r \geq 1} \lambda_r c_{rj} = 0$. We write $Z = e^{1-\lambda_0}$ to obtain a simple formal solution:

$$p_j = \frac{1}{Z} \exp(-\sum_{r \geq 1} \lambda_r c_{rj}) . \qquad (6)$$

We can always eliminate the normalization $(r = 0)$ constraint: taking

$$Z = \sum_j \exp(\sum_{r \geq 1} -\lambda_r c_{rj}) \qquad (7)$$

guarantees that $\sum p(x) = 1$.

Now we need to adjust the λ's so that the constraints are satisfied. But even before we do this, we can see the form of the solution: the probability distribution has an exponential form, and the constraints enter into the exponent. Adjusting the λ's can be easy, tricky, or even intractable, depending on the nature of the linear constraints.

All of the above works fine when the distributions are continuous; the "fine print" concerning the $\log(\Delta x)$ -term doesn't matter since it just leads to an additive offset on all entropies. In the continuous case, the constraint descriptors and the probabilities depend on a continuous variable *x* rather than a discrete index *j*, and eq. (6) becomes

$$p(x) = \frac{1}{Z}\exp(-\sum_{r\geq 1}\lambda_r c_r(x)), \tag{8}$$

and *Z* is determined by $\int_X \exp(-\sum_{r\geq 1}\lambda_r c_r(x))dx = 1$

A nice feature of this setup (generic to maximum-entropy problems with linear constraints, but not to all Lagrange Multiplier problems) is that the constraint equations can be expressed simply in terms of derivatives of *Z*. Using the discrete formulation (7):

$\frac{\partial Z}{\partial \lambda_r} = \frac{\partial}{\partial \lambda_r}\left(\sum_j \exp(-\sum_r \lambda_r c_{rj})\right) = -\sum_j c_{rj}\exp(-\sum_r \lambda_r c_{rj}) = -Z\sum_j c_{rj}p_j = -Zb_r$, where we've used eq. (6) for p_j. Since $\frac{\partial \ln G}{\partial u} = \frac{1}{G}\frac{\partial G}{\partial u}$, it follows (both for the discrete formulation and the continuous one) that

$$-\frac{\partial \ln Z}{\partial \lambda_r} = b_r, \tag{9}$$

a useful and compact expression of the constraints.

The equations (9) are sometimes called the "conjugate" problem for the original extremization.

One cannot help but mention the connection with statistical mechanics. The λ's correspond to energies, and *Z* is the "partition function".

Basic examples

No constraints

The simplest (possibly trivial) example is to find the maximum entropy distribution on an interval $[a_0, a_1]$ without any constraints (other than normalization). The exponential term in eq(8) vanishes, so $p(x) = 1/Z$. The normalization condition is that $1 = \int_{a_0}^{a_1} p(x)dx = \frac{a_1 - a_0}{Z}$, so

$Z = a_1 - a_0$. Not surprisingly, we see that if there are no other constraints, a maximum-entropy distribution is uniform.

Constrain the mean

The next example is to find a maximum-entropy distribution on the interval $[0,\infty]$, for which the mean is specified: $\int_0^\infty xp(x)dx = \mu$. We apply the above machinery, associating this constraint with a Lagrange Multiplier λ_1 and $c_1(x) = x$. Eq. (8) yields $p(x) = \frac{1}{Z}e^{-\lambda_1 x}$. Our constraints are

$1 = \int_0^\infty \frac{1}{Z}e^{-\lambda_1 x}dx$ (normalization) and

$\mu = \int_0^\infty xp(x)dx = \int_0^\infty \frac{1}{Z}xe^{-\lambda_1 x}dx$ (the condition on the mean).

The definite integrals specified by the constraints are undefined unless $\lambda_1 > 0$. Provided that this is the case, we find

$\int_0^\infty \frac{1}{Z}e^{-\lambda_1 x}dx = -\frac{1}{Z}\frac{e^{-\lambda_1 x}}{\lambda_1}\Big|_0^\infty = \frac{1}{Z\lambda_1}$, which must be equal to 1, and

$\int_0^\infty xp(x)dx = \int_0^\infty \frac{1}{Z}xe^{-\lambda_1 x}dx = \frac{1}{Z}\left(-\frac{xe^{-\lambda_1 x}}{\lambda_1} - \frac{e^{-\lambda_1 x}}{{\lambda_1}^2}\right)\Big|_0^\infty = \frac{1}{Z{\lambda_1}^2}$ which must be equal to μ, The first condition (normalization) implies that $Z = 1/\lambda_1$, and second implies that $1/\lambda_1 = \mu$. Thus, $p(x) = \mu e^{-x/\mu}$.

Note that had we sought to solve this problem on the full line $[-\infty,\infty]$, the machinery would have told us that there is no solution. This is correct … one can make the entropy as large as desired by making the distribution as "thin" as desired (e.g., allow $a_0 \to -\infty$ and $a_1 \to \infty$ in the previous example).

Constrain the mean and variance

The next example is to find a maximum-entropy distribution on the interval $[-\infty,\infty]$ which the mean is specified, and the variance is too: $\int_{-\infty}^\infty xp(x)dx = \mu$ and $\int_{-\infty}^\infty (x-\mu)^2 p(x)dx = V$. This yields the formal solution

$p(x) = \frac{1}{Z}e^{-\lambda_1 x - \lambda_2 (x-\mu)^2}$. The definite integrals specified by the constraints are only defined if $\lambda_2 > 0$. We use eq. (9), so that the only calculus needed is to calculate Z:

$$Z = \int_{-\infty}^{\infty} e^{-\lambda_1 x - \lambda_2 (x-\mu)^2} dx = \sqrt{\frac{\pi}{\lambda_2}} \exp\left(\frac{{\lambda_1}^2}{4\lambda_2} - \mu\lambda_1 \right)$$ (a standard definite integral).

Eq. (9) leads to the constraint equations:

$$-\frac{\partial \ln Z}{\partial \lambda_1} = -\frac{\lambda_1}{2\lambda_2} + \mu = b_1$$ (for the mean) and $$-\frac{\partial \ln Z}{\partial \lambda_2} = \frac{1}{2\lambda_2} + \frac{{\lambda_1}^2}{4{\lambda_2}^2} = b_2$$ (for the variance).

Since our constraints are that $b_1 = \mu$ and $b_2 = V$, the constraint equations are satisfied by $\lambda_1 = 0$ and $\lambda_2 = 1/2V$, yielding $Z = \sqrt{2\pi V}$ and

$$p(x) = \frac{1}{\sqrt{2\pi V}} e^{-(x-\mu)^2/2V} . \qquad (10)$$

Eq. (10) says that the maximum entropy distribution with a specified mean and variance is a Gaussian.

We've derived this important fact in a systematic fashion, but we also could have "solved" the Lagrange multiplier problem by guessing that the Gaussian plays this role, and then verifying that eq. (10) has the correct form – an exponential of the constraints. (Convexity guarantees that there is only one solution.) The latter approach proves useful in more complex maximum-entropy problems.

Two brief notes:

The "guess" method provides another way to show that for a multivariate distribution $p(x, y, ...z)$ in which only the marginals are constrained, maximum entropy is achieved for a product distribution $p(x, y, ..., z) = p(x)p(y) \cdot ... \cdot p(z)$, since the latter has the correct exponential form: $p(x, y, ..., z) = \frac{1}{Z} \exp(-\lambda_x - \lambda_y - ... - \lambda_z)$.

We also note that once one adds constraints beyond the second moment, closed-form solutions are in general not possible.

A more complex example: time series with constrained covariances

A more complex example is to determine the maximum-entropy distribution (ensemble) of time series, in which the autocovariance is specified. (We also specify that the mean is zero.) Such ensembles formalize what one has learned about a process by measuring its autocorrelation (or power spectrum), and can therefore be used as a starting point for statistical tests.

The strategy here extends with little difficulty to multichannel time series for which the cross-covariances (or cross-spectra) have been measured.

At first glance, this might seem to be a difficult problem: each constraint – the value of an autocorrelation at a specific lag τ – is a constraint on a sum such as $\sum_t s(t)s(t+\tau)$, and there is one constraint for each τ. So, while we could write a formal solution – something like $p(x) = \frac{1}{Z}\exp\left(-\sum_\tau \lambda_\tau \left(\sum_t s(t)s(t-\tau)\right)\right)$, it would appear to be very challenging to carry out the multidimensional integral for *Z* or to solve the constraint equations.

Not surprisingly, transforming to the frequency domain simplifies things greatly. We've seen previously that specifying the autocovariance is equivalent to specifying the spectrum. Our coordinates are now the Fourier estimates (over some arbitrarily long time interval.) We want to constrain the expected magnitude of each Fouirer estimate, but we don't need to constrain their covariances … since for the maximum-entropy distribution, these must be independent.

We can then consider the distribution of each Fourier estimate independently. Its mean is constrained to be 0 (time-translation symmetry), and its variance is determined by the Fourier transform of the autocovariance. So each Fourier estimate is distributed as a Gaussian, of zero mean and known variance. The joint distribution of the Fourier estimates is then the product of these distributions.

The resulting process is known as a Gaussian noise.

To see how time samples are distributed, we transform back to the time domain. Each time point is a Fourier integral $s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{s}(\omega)e^{+i\omega t}d\omega$. Since this is a sum of independent Gaussian-distributed quantities, it is Gaussian-distributed. Covariances at different time points will (by construction) be equal to the constrained values. Importantly, even if the original constraints did not exceed some longest lag τ_m, the maximum-entropy process will have correlations at lags greater than τ_m.

Sometimes "phase-scrambling" is used as a shortcut to generate surrogate datasets that share the same spectrum (or autocorrelation) as the data, but are otherwise of maximum-entropy. That is, the data are Fourier-transformed; random values are assigned to the phases of the Fourier components, and then the transform is inverted to recover new time series. The reason that this works is that each Fourier component (with respect to the entire data length) can be regarded as an independent quantity, and adding them up results in a Gaussian quantity (via the central limit theorem). However, one needs to be cautious -- it is only an approximate construction, and Fourier components of the surrogates will (by construction) always have the same amplitude as that of the original data; in a true Gaussian ensemble, these would be distributed in a Gaussian fashion.