Methods of Data Analysis
Introduction to information theoretic quantities

Week 4

1 Motivation

Joint distributions provide a full mathematical account of dependencies between the system’s degrees of freedom. Up to now, we focused on probabilistic models as a step towards formalizing the inference of parameters given the data, often using examples where the data is reasonably simple; the non-trivial nature of the problem emerged from how parameters mapped to observations. Conversely, the data itself could be very high dimensional, in which case inferring most probabilistic models can be difficult when the data is scarce. Such inferences will be the topic of the next lecture. Here, we return to the idea that joint distributions contain all possible statistical dependencies between the variables. How can these dependencies be formalized and computed from the joint distribution? Which distributions are “simple” and which ones “complicated”? If we find good candidate measures of dependence (which are often low-dimensional quantities) defined on the joint distribution, can we construct direct estimators for them rather than first using the data to infer the distribution itself, which might be prohibitively hard?

As an example, consider the commonly used measure of statistical dependence between two variables, $x$ and $y$, the (linear) correlation coefficient, also known as Pearson Correlation $r$:

$$PC = r(x, y) = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sigma_x \sigma_y}, \quad (1)$$

where $\langle \cdot \rangle$ are empirical averages, and $\sigma_x, \sigma_y$ are the standard deviations of $x$ and $y$, respectively. PC is bounded within $[-1, 1]$, with $|r| = 1$ indicating perfect correlation or anti-correlation. It is easy, however, to find statistical dependencies between two variables that are not captured by PC. For instance, in a spherically symmetric distribution where $P(x, y) = \delta(1 - x^2 - y^2)$ the two variables are perfectly related (since they live on the unit circle), but for such data PC will be exactly equal to zero. Is there a general measure of pairwise statistical dependence that will detect such relations? How should it be computed over samples drawn from $P$?

Fundamental in approaching these questions are information-theoretic quantities. We will see that information theory defines a unique measure of statistical dependence, the mutual information, which is zero if and only if a pair of random variables is statistically independent. This measure is universal, in that it can be applied to any pair of variables over which a joint distribution can be defined. It also generalizes to statistical dependence between more than two variables, and related quantities are able to systematically isolate effects of intrinsically pairwise, triplet, quadruplet, etc. interactions between random variables. On the downside, estimating information-theoretic quantities from limited data is notoriously hard, because the naive estimator is strongly biased, and schemes for removing the bias do not universally work.
Thinking about (pairwise) covariance has given rise to useful dimensionality reduction techniques and transformations of the data (most directly, to Principal Component Analysis and whitening, which linearly rotates and scales the data such that the transformed variables have unit covariance matrix). Similarly, thinking about mutual information gave rise to more powerful ideas where the data is linearly transformed such that the resulting variables are as statistically independent as possible (in independent component analysis, or ICA). Many likelihood-based approaches and quantities known from Bayesian inference are intimately linked to information-theoretic quantities; while these connections are theoretically very interesting, they are beyond the scope of this lecture.

2 Goals

- Define Shannon entropy.
- Define KL and JS divergences, mutual information, multi-information, redundancy / synergy, information decomposition.
- Mutual information for Gaussian variables and Gaussian processes.
- Estimation of entropy and mutual information from data.

3 Data

- A subset of gene expression studies from Ref [2]: a matrix of 200 genes by 173 conditions, where each value represents log fold-change in a gene expression level (microarray) in yeast, under a specific condition relative to the control condition.
- A draw of $2^{16}$ 6-bit samples from an unknown distribution, saved in mda4.data.txt.

4 Entropy

Shannon considered "information-producing" sources, which he modeled as stationary stochastic processes that are emitting symbols in a particular sequence. As an example (that he was also interested in) you can consider sequences of characters that represent the written English language. His wonderfully readable classic work [1] asks about the "rate" at which the information is produced by such a source. Here we outline the basic reasoning. Let the information source emit symbols $i$ with probabilities $P_i$; we are assumed to know this distribution. What we don’t know, however, is the next generated symbol. How much choice is there really in the selection of the symbol, or how uncertain are we of the outcome? Shannon looked for a measure $H$ to address these questions which would satisfy three conditions:

- $H$ should be continuous in $P_i$.
- If all $P_i$ are equal, $P_i = 1/N$ where $N$ is the number of possible symbols that the source can emit, then $H$ should be monotonically increasing function of $N$. This is because with equally likely events, there is more choice (or uncertainty) when there are more possible events.
If a choice can be broken down into two successive choices, the original $H$ should be a weighted sum of the individual values of $H$. Think of a game of twenty questions to see why this requirement makes sense.

Shannon then proceeds to show that the only $H$ satisfying these properties is of the form

$$H = -K \sum_{i=1}^{N} P_i \log P_i.$$ 

The related “entropy rate” of the source is then $H_f$, where $f$ is the frequency of symbol emission. This can be generalized to Markovian sources with internal states, etc. An important property of this definition is the asymptotic equipartition (or typicality), which corresponds to the following assertion. If we consider sequences composed of $M$ symbols, for $M$ sufficiently large, all sequences can be split into two disjoint sets; the total weight of sequences in one set can be made arbitrarily small, whereas in the second set (“typical set”) we find sequences with probability $p$ such that

$$\log \frac{1}{p} \approx MH$$

to within desired accuracy.

Choice of $K$ is obviously related to the base of the logarithm, with the standard Shannon entropy $S$ commonly using base 2 (in which case the result is given in “bits”), and the standard physics using natural log and being expressed in units of the Boltzmann constant, $k_B$. Here we will write $S$ for standard Shannon entropy:

$$S[P] = - \sum_x P(x) \log_2 P(x). \quad (2)$$

For a binary 0/1 variable that can take on the value 1 with probability $p$, the entropy $S_2$ (written out explicitly as a function of parameter $p$) is

$$S_2(p) = -p \log_2 p - (1-p) \log_2 (1-p). \quad (3)$$

This function is symmetric around $p = 0.5$, where it reaches 1 bit, and then decays to zero for $p = 0$ or $p = 1$, consistent with the expectation that at those extremes, the outcome of the stochastic process is fully predictable and that thus the entropy, as a measure of uncertainty, should vanish. This is true in general: the entropy of a discrete distribution is zero if and only if for all $P$, one of them is one while the rest are zero; otherwise entropy is positive. The entropy is maximal for the uniform distribution, in which case it is equal to $\log_2 N$ bits.

One more property of the entropy will be important. Consider a joint distribution $P(x,y)$ over two events. Then, one can show that

$$S[P(x,y)] \leq S[P_x(x)] + S[P_y(y)], \quad (4)$$

where $P_x(x)$ is the marginal distribution, $P_x(x) = \sum_y P(x,y)$, and similarly for $P_y(y)$. So the entropy of the joint distribution is always smaller or equal to the sum of the entropies of the two marginal distributions. Intuitively, any type of statistical dependency between $x$ and $y$ will decrease the joint entropy, but not the marginals; the equality will hold when $x$ and $y$ are statistically independent and the joint distribution factorizes, $P(x,y) = P_x(x)P_y(y)$.

For continuous distributions,

$$S[p] = - \int dx \ p(x) \log_2 p(x). \quad (5)$$

In this case, one has to be careful because of the issues of measure (and for physicists, units). For instance, for a Gaussian, $G(x) = (2\pi \sigma^2)^{-1/2} \exp(-x^2/\sigma^2)$, the entropy is $S[G] = \frac{1}{2} \log_2(2\pi e \sigma^2)$, which has a nonintuitive behavior as $\sigma \to 0$. To appreciate the source of weird behavior of entropy for continuous distributions, we can approximate a continuous distribution with a sequence of discrete distributions defined on bins of size $\Delta$. A simple calculation then shows that
the entropy of the continuous distribution, \( p(x) \), and its discrete approximation, \( P_i = p(x_i) \Delta \) (such that \( \sum_i P_i = 1 \)), are related as \( \Delta \rightarrow 0 \) as

\[
S[p] = S[P] + \log_2 \Delta.
\] (6)

For a continuous distribution, a better suited quantity is the relative entropy or the KL-divergence, see below. Nevertheless, entropy differences are well-behaved because the log \( \Delta \) terms subtract out and there are no divergences in bin size (it is interesting to note that this is similar to classical thermodynamics, where only entropy differences make sense). Mutual information and other information measures are often expressible as entropy differences, and thus they can safely be used on discrete or continuous distributions. Interpretation-wise, entropy of a distribution is a measure of the diversity of values the random variables can take.

Another way to view the entropy is as a measure of average surprise; \(-\log p_i\) is the surprise of symbol \( i \) (that is, rare symbols are more surprising when seen). Often, the entropy of a distribution is mistakenly referred to as the “information” in the distribution, which causes a lot of confusion. As we will see below, the proper definition of the (mutual) information is always on a joint distribution, since one random variable can hold information about another variable, but in itself, a random variable does not contain information. This confusion arises because entropy bounds the information, and so the entropy of some random variable \( x \), \( S[P(x)] \), is an upper bound to the mutual information \( x \) can provide about any other variable.

5 Divergence measures

Divergence measures are functionals of pairs of distributions, and intuitively correspond to the notion of dissimilarity between the distributions. The most important divergence measure between a pair of distributions is the Kullback-Leibler divergence, defined as

\[
D_{KL}(p||q) = \int dx \, p(x) \log_2 \frac{p(x)}{q(x)}.
\] (7)

This is a non-negative quantity, measured in bits, and also known as relative entropy. It can be understood as an entropy in \( p \), relative to some prior distribution \( q \). KL-divergence is well-behaved for continuous distributions. It is often used as a distance measure between distributions \( p \) and \( q \), since \( D_{KL} = 0 \) if and only if \( p = q \). Strictly speaking, the distance interpretation is incorrect, since the KL divergence is neither symmetric nor does it satisfy the triangle inequality. Nevertheless, \( D_{KL} \) often comes up in theoretical derivations, and can actually be used as a basic quantity from which other information theoretic quantities can be derived. Interpretation-wise, \( D_{KL} \) is the number of additional bits needed to code a random variable \( x \), if \( x \) is really drawn from \( p \), but the code was designed for \( x \) drawn from \( q \). Jensen-Shannon divergence is a normalized (bounded between 0 and 1), symmetric “distance” measure between two distributions, defined as \( D_{JS}(p, q) = 0.5(D_{KL}(p||r) + D_{KL}(q||r)) \), where \( r = 0.5(p + q) \).

An interesting connection between likelihood and information-theoretic quantities arises when we consider inference. Suppose we have a parametrized probabilistic model, \( P(y|\theta) \), where \( \theta \) are the parameters and \( y \) are the observable outcomes. We observe data \( D = \{y_t\} \), with \( t = 1, \ldots, T \) samples. We can represent the samples as being drawn from an empirical distribution, \( \hat{P}(y) = \frac{1}{T} \sum_t \delta(y - y_t) \), where \( \delta \) is the Dirac-delta function. One possible way to formalize the inference would be to look for model parameters that minimize the KL divergence between the model distribution and the empirical data distribution, that is:

\[
D_{KL}(\hat{P}(y)||P(y|\theta)) = \int dy \, \hat{P}(y) \log_2 \frac{\hat{P}(y)}{P(y|\theta)}.
\] (8)
By plugging into the above equation the formula for the empirical distribution over data, it is easy to show that the $D_{KL}$ can be written as

$$D_{KL}(\hat{P}(y)||P(y|\theta)) = -\frac{1}{T} \sum_i \log P(y_i|\theta) - S[\hat{P}(y)].$$

(9)

The second term is the entropy of the empirical distribution, which is independent of parameters, $\theta$. But the first term is nothing else than the (log)likelihood, showing that minimization of $D_{KL}$ between the model and the data distribution for inference is consistent with maximum likelihood inference.

6 Mutual information

Given a joint distribution over two random variables $x$ and $y$, the mutual information between $x$ and $y$ is

$$I(x; y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P_x(x)P_y(y)},$$

(10)

Information $I$ is a scalar, a functional of $P(x,y)$ (despite the traditional notation, it is not a function of $x$ or $y$). $I$ is non-negative, in bits, and is a measure of the total amount of statistical dependence between $x$ and $y$ ($I = 0$ only when they are independent, i.e., $P(x,y) = P_x(x)P_y(y)$).

It is very instructive to rewrite Eq. (10) in various equivalent forms. First, note that

$$I(x; y) = S[P_x(x)] + S[P_y(y)] - S[P(x,y)],$$

(11)

so the mutual information is precisely the difference in entropy between marginal distributions and the joint, which (as we stated above) arises due to any statistical dependency between $x$ and $y$. Alternatively, you can convince yourself that

$$I(x; y) = D_{KL}(P(x,y)||P_x(x)P_y(y)),$$

(12)

that is, the mutual information is the divergence (“distance”) between the joint and the factorial distributions. Mutual information can also be rewritten as

$$I(x; y) = S[P_y(y)] - \langle S[P(y|x)] \rangle_{P_x(x)},$$

(13)

i.e., the average reduction in the uncertainty about the $y$ (from its total entropy, first term) when $x$ is known (by noise entropy, the second term). If given $x$, $y$ is unambiguously determined, for instance, by a functional map $y = f(x)$, then $P(y|x) = \delta(y - f(x))$, whose entropy is zero, and the mutual information between $x$ and $y$ will be equal to the entropy of the outcomes, $y$; i.e., $I(x;y) = S[P_y(y)]$. This can be smaller than the entropy of the inputs $x$, $S[P_x(x)]$, if the functional map is not bijective. Information is in general upper-bounded by the entropy, because the second term in Eq. (13) is always non-negative.

$I$ is symmetric, $I(x; y) = I(y; x)$, as is obvious from Eq. (10). Therefore, Eq. (13) can be rewritten with the roles of $x$ and $y$ exchanged, showing that information is bounded from above by the smallest of the two marginal entropies.

Information is also reparametrization invariant, so $I(x; y) = I(f_x(x); f_y(y))$, where $f_x$ and $f_y$ are invertible functions of their arguments. This follows simply from the fact that the information is a functional of the distributions which, under change of variables, must transform to keep the integrals invariant. This is very important for data applications, since many correlation measures

5
Figure 1: Red points represent 1000 jointly Gaussian samples with specified Pearson Correlation coefficient (PC). The corresponding mutual information is computed analytically from the coefficient and shown in the plot title in bits, along with the PC. If the scattered points represented experimental data, then very high measurement precision would be needed to claim information values of significantly more than 1 bit (since experimental noise would degrade the information).

don’t share this invariance (e.g., if you compute correlation coefficient between $x$ and $y$, or $\log(x)$ and $\log(y)$, you get a different answer).

Another useful result is the so-called Data Processing Inequality. Consider the following dependence between random variables, $X \rightarrow Y \rightarrow Z$, where $X \rightarrow Y$ means that $Y$ conditionally depends on $X$ (but that $Z$ is conditionally independent of $X$ given $Y$). Then it follows that $I(X;Z) \leq I(X;Y)$, i.e., information can only be lost, not created, by successively chaining together probabilistic processes. Such inequalities can be used to build (or rule out) candidate dependency graphs between the variables from data.

If $x$ and $y$ are jointly Gaussian with a correlation coefficient $c$, then the mutual information $I(x;y) = -0.5\log_2(1-c^2)$. If $x$ and $y$ are vectors whose distribution is jointly Gaussian, then the information is

$$I(x;y) = \frac{1}{2} \log_2 \left( \frac{\det C_{xx} \det C_{yy}}{\det C} \right), \tag{14}$$

where $C_{xx}$ and $C_{yy}$ are the covariance matrices of the marginal distributions $P_x(x)$ and $P_y(y)$, respectively; and $C$ is the covariance matrix of the joint Gaussian. Note that this formula can also be used to estimate the mutual information between two jointly Gaussian time series.

Shannon has shown that $I$ is the only (up to the unit of measure) quantity consistent with certain basic postulates that an information measure should have (e.g., such as that independent pieces of information add). Interpretation-wise, $2^{I(x;y)}$ is the approximate number of distinguishable levels or states in $y$ accessible by changing $x$, given noise. To gain some intuition, compare the correlation coefficients and the mutual information values for jointly Gaussian random variables in Fig. 1.

While $I$ provides an assumption-free measure of statistical dependence, this measure lacks some of the appealing properties of the correlation coefficient. First, it is not normalized. This can be addressed by dividing $I$ with the (minimum) entropy of the two marginal distributions, in which case $I/S \in [0, 1]$. On the other hand, the absolute unit of information actually has independent meaning which is relevant for some applications. Generally, since estimating $I$ reliably can be difficult, we should carefully study if our application calls for the true, unbiased value of information, whose absolute value is important; or if we are treating naive estimates of $I$ purely as useful statistics of the data that may have comparative, but not absolute, relevance (e.g., that $I$ between a pair of variables is higher than between a certain other pair). Second, $I$ says nothing about the kind of statistical dependence, just its “magnitude”. This is in contrast...
to PC, which directly relates to the goodness-of-fit of a linear model to data. Information thus makes explicit the separation between detecting, and modeling, probabilistic dependencies between the variables.

7 Related information-theoretic quantities

Multi-information and connected information quantities. Consider a joint distribution over several variables, \( P(x_1, x_2, \ldots, x_N) \). Multi-information measures the total amount of statistical correlation between all the variables, and is defined as

\[
I = S[P(x_1)P(x_2) \ldots P(x_N)] - S[P(x_1, \ldots, x_N)],
\]

i.e., the difference in entropies of the joint and the factorial distributions; for two variables, this is the mutual information. There exists a unique decomposition of the total multi-information into positive contributions, \( I = I^{(2)} + I^{(3)} + \ldots + I^{(N)} \), where \( I^{(k)} \) is called the connected information of order \( k \), and it measures the fraction of total correlations contributed by the irreducible interactions between exactly \( k \) variables. The construction of this decomposition is difficult in practice and rests on constructing a series of approximating distributions to the joint (called the maximum entropy distributions consistent with correlation of up to order \( k \), which we will discuss in the next lecture). This is a principled way of assessing at which order the variables interact (i.e. if \( I^{(2)} \) accounts for a large fraction of total \( I \), pairwise interactions between variables are sufficient to explain all the statistical structure in the joint distribution). For details on this decomposition, see Ref [4].

Redundancy and synergy. Consider variables \( x \) and \( y \) that both carry some information about the input, \( s \). For a concrete example, let \( x \) and \( y \) be responses of two neurons that both respond to the common stimulus \( s \). Given the quantities we already defined, we can ask how much information each of the neurons carries about the stimulus, and that is quantified by the mutual information, \( I(x; s) \) and \( I(y; s) \). From the two responses, we can form a new quantity, \( z = [x, y] \), and ask how much information do both responses combinatorially encode about \( s \), that is, what is \( I(x, y; s) = I(z; s) \). An interesting quantity is redundancy (sometimes expressed fractionally, normalized by the information, \( I(x, y; s) \)):

\[
R(x, y; s) = I(x; s) + I(y; s) - I(x, y; s).
\]

If \( R > 0 \), the neurons are redundant, that is, they encode overlapping information such that the sum of their individual informations is larger than the combinatorial response they carry. Suppose the response of the two neurons would be exactly identical. Then \( I(x, y; s) = I(x; s) = I(y; s) \) and thus \( R = I(x; s) > 0 \). In contrast, \( R < 0 \) signifies synergistic (combinatorial) coding.

8 Entropy and information estimation

The plug-in (or naive, or maximum-likelihood) estimator of entropy is

\[
\hat{S}[P] = -\sum_i \hat{P}_i \log_2 \hat{P}_i,
\]

where \( \hat{P}_i \) is the sample estimate of the probability of the \( i \)-th event.
where $\hat{P}_i = n_i / \sum_i n_i$ is the empirical frequency at which state $i$ occurs in a sample of $\sum_i n_i$ total observations. This estimator is always negatively biased. To see that, consider true probabilities, $P_i$, and write $\hat{P}_i = P_i + \delta P_i$, where the error $\delta P_i$ is the small sample noise in the empirical estimate of probability. We can then write

$$\hat{S} = -\sum \hat{P}_i \log_2 \hat{P}_i$$

(18)

$$= -\sum \hat{P}_i \log_2 \frac{\hat{P}_i}{P_i}$$

(19)

$$= -D_{KL}(\hat{P}||P) - \sum (P_i + \delta P_i) \log_2 P_i$$

(20)

We can then take the expectation over data draws, such that $\langle \delta P_i \rangle = 0$ to find

$$\langle \hat{S} \rangle = S - \langle D_{KL}(\hat{P}||P) \rangle.$$

(21)

Since the Kullback-Leibler divergence is always non-negative, this equation implies that the naive estimate always underestimates the true entropy. In most real-life situations, this bias (rather than the variance in the estimate because of the small sample effect) is the dominant problem that estimators have to deal with, as shown on a simple example in Fig. 2. In fact, often the number of total observations, $N = \sum_{i=1}^{K} n_i$ might be much smaller even that the total number of possible states $K$ of the system (it would seem that in such a situation, no estimator can ever do the job, but this really depends on the underlying probability distribution).

The “direct” estimate is based on the observation that the bias term scales inversely with $N$, the total sample size, so that $\langle \hat{S} \rangle = S + A/N + \cdots$, where $S$ is the true entropy. As a result, one can take many random subsets of fractions of data, e.g., $M(f_1)$ random choices of $f_1 N$ observations, where $f_1$ is the fraction between 0 and 1, $M(f_2)$ random choices of $f_2 N$ observations etc. For each random choice of observations, one estimates the empirical distribution, and the resulting naive estimate of entropy, to get $\langle \hat{S}(f) \rangle$, the average (over random selections of observations) naive estimate at a given data fraction. This can be plotted versus the inverse data size, $1/(fN)$, and linearly extrapolated to infinite data size (i.e., towards zero), as shown in Fig. 3. This extrapolation removes the bias and extends the range of validity of the estimation beyond the point where naive estimator would already give strongly unreliable results. For a specific example relating to neural code, see Ref [7] which introduced the direct method.

Information, as a difference of two entropies, similarly has an inverse-$N$ bias (but its sign is arbitrary), and can be debiased in an analogous way. A direct estimator for the information has been presented in Ref [11]; follow this reference to implement the mutual information estimator for continuous data. The key point here is that continuous data need to be discretized, but the “proper” number of levels is a priori unknown. More levels allows for capturing finer features in the joint distribution, but given the finite data sample, also leads to increasingly large bias problems. Given the data, the task is thus to select the maximal possible binning level such that the debiasing is still reliable, while information is not lost due to discretization.

Note that there exist other more advanced estimators both for discrete probabilities [8] and for binless estimation over continuous probability distributions [6].

9 Study literature

- For formal definitions of information theoretic quantities, see a classic book on information theory, Ref [3].
Figure 2: The bias of the naive entropy estimator. The assumed true distribution is uniform over $2^{10}$ bins, i.e., $S_{\text{true}} = 10$ bits. We draw $N$ samples 100 times, where $\log_{10} N = \{2, 2.5, 3, 3.5, \ldots, 5.5\}$. From the samples, we construct the empirical distribution, $\hat{P}$ (see text) and estimate its entropy. Top: the distributions of naive entropy over 100 replicates using different number of samples, $N$. Leftmost distribution corresponds to $N = 100$; successive distributions correspond to more samples, converging to the true estimate of 10 bits when $N$ is sufficient. Bottom: the mean information and std (error bars) over replicates as a function of $N$. Black curve = naive estimator. Red curve = de-biased “direct” estimator.
Figure 3: Extrapolations in inverse sample size for the direct estimator. The assumed true distribution is uniform over $2^{10}$ bins, i.e., $S_{\text{true}} = 10$ bits. Each red line (with black dots) represents a debiasing run for different values of $N$, as in Fig. 2; smaller $N$ start at lower information values and show steeper extrapolation slopes. At each value of $N$, data is subsampled 100 times at different fractions without replacement (x axis), and the average of the naive estimators is plotted in black; these naive estimates are then extrapolated (here quadratically) to the infinite sample limit. The de-biased estimate is thus the y-axis intercept of the red extrapolation lines.

- Direct entropy estimation, Ref [7]; Nemenman-Shafee-Bialek (NSB), one of the best performing entropy estimators for discrete distributions, Ref [8]; and the recent advance on it, Ref [9]. Entropy estimation without binning for continuous distributions, Ref [5], and kNN (k-Nearest-Neighbor estimator) [6]. An interesting bound-on-entropy estimator, but the paper contains a great overview of various estimators, Ref [10].

10 Homework

1. Implement a direct estimator for entropy for a discrete distribution (i.e., debias the estimate by extrapolating to infinite sample size). Use samples from data.mat (which you are going to use also in a subsequent problem). Randomly choose $2^3, 2^4, 2^5, \ldots, 2^{16}$ samples from the data, and estimate (i) the entropy using your direct estimator; (ii) the entropy in a naive way, by counting the frequency of patterns and plugging in into the entropy formula. When you run the direct estimator which extrapolates naive entropy to infinite sample size, show the extrapolation plot for the run with a total of $2^{10}$ samples and for the run with a total of $2^{16}$ samples (i.e., plot the estimated naive entropy as a function of $1/(\text{sample size})$, and the best fit linear extrapolation extending to 0 on the x-axis, to find the y-intercept). Do you see the expected scaling behavior for the small sample correction? Show the entropy for both methods as a function of the number of samples. Optional: download and use the NSB estimator ([8], or its improved version, [9]) and compare its performance with the direct estimation (if you work with discrete distributions, these estimators may anyway come in handy).

2. Implement the direct estimator for information between continuous variables, following [11]. First, to check that everything works properly, let’s generate synthetic data for which
we know the correct answer: pairs of jointly Gaussian distributed random variables with given correlation coefficients, equal variance, and zero mean. Generate random samples for 20 such pairs with correlation coefficients $c$ going uniformly from −0.95 to 0.95 in equal increments. Think carefully and explain how you generate pairs of normally distributed random variables with a desired correlation coefficient. *(Optional) Do you know how to generate many jointly gaussian distributed variables with the prescribed covariance matrix?* For each pair, draw 10, 100, 1000, 10000 samples, estimate the mutual information $I$ using the direct estimator. Plot the estimated $I$ vs the correlation coefficient for all pairs and for all 4 sample sizes (different sample sizes using different colors / plot symbols), and compare on the same plot with the analytical curve $I = -0.5 \log_2(1 - c^2)$. Do you have an intuitive explanation what it means that at finite correlation coefficient you can have a diverging mutual information?

3. Next, use the direct estimator to analyze Gasch data on gene expression. If your algorithm is fast enough, compute the mutual information for all pairs of genes, otherwise, randomly select 1000 pairs and estimate the information for them. Let your estimator output both the corrected (debiased) information value as well as the naive value, and make a scatter-plot of the corrected vs naive value. How big is the small sample bias for this dataset? What is the typical (median / mean) value of information across the analyzed pairs? What is the maximal value that you observe? Beyond factors that are intrinsic to the biological system under study, what factors (that have nothing intrinsic to do with the cell) influence the magnitude of the observed numbers? Plot the information values for all pairs on y axis, vs the correlation coefficient for those pairs, on the x axis, and compare with the predicted curve for jointly Gaussian variables. Find several pairs that deviate the most for the curve: (i) pairs that maximal information given a very small correlation coefficient, e.g., $|c| < 0.05$; (ii) pairs that have a very small amount of mutual information yet have a large correlation coefficient, e.g., $|c| > 0.6$. For several of these pairs $(g_i, g_j)$, scatter plot the data of $g_i$ vs $g_j$ (each point is a joint expression level in a given condition). Can you propose a few suggestions about the data that explain the particular deviations you see?

4. Analyze the 6-bit binary patterns in *data.mat* which were generated by drawing IID from an unknown distribution. Figure out the rule that generated the patterns and, if you can, write down the generating distribution. Hint: try to analyze the statistical structure order-by-order. You can start by computing the pairwise correlation coefficient between the variables, and the mutual information between all pairs. Suggest and carry out further analyses to determine the statistical structure in the distribution.

References


