Estimation of Entropy Combinations

Kalle Rutanen

March 11, 2010

1 Abstract

This paper introduces *entropy combinations* as a general subclass of Kullback-Leibler divergences for which an efficient estimator can be derived. Such an estimator is derived by a natural extension of Alexander Kraskov's mutual information estimator [1]. This allows a single algorithm to be used for all such types of computations. The estimator is based on combining estimates from a Kozachenko-Leonenko estimator for Shannon differential entropy, and thus this estimator is also derived. Finally, examples are given for mutual information, partial mutual information, total correlation, transfer entropy, and partial transfer entropy.

2 Notation

Let $\{(M_1, d_1), \ldots, (M_m, d_m)\}$ be a set of metric spaces, where $M_i = \mathbb{R}^{b_i}$, and $d_i : M_i^2 \to \mathbb{R}$ are metrics. If $I = \{i_1, \ldots, i_{|I|}\} \subset [1, m]$, then $M_I = M_{i_1} \times \cdots \times M_{i_{|I|}}$. Define a metric in M_I by $d_I(x, y) = \max(d_{i_1}(x_1, y_1), \ldots, d_{i_{|I|}}(x_{|I|}, y_{|I|}))$. Thus (M_I, d_I) is also a metric space. For brevity, denote $M = M_{[1,m]}$ and $d = d_{[1,m]}$. We shall call M the *joint space*, and the M_i the *marginal spaces of* M.

An ϵ-radius ball around $x \in M_i$ is defined by $B_i(x, \epsilon) = \{y \in M_i : d_i(x, y) \leq \epsilon\}$. The Lebesgue measure of a set S is denoted by $m(S)$. The characteristic function of a set S is denoted by χ_S. The gamma function is

denoted by Γ. The beta function β and it's derivative are defined by:

$$\begin{aligned}
\beta(r,s) &= \int_0^1 t^{r-1}(1-t)^{s-1}dt \\
&= \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} \\
\beta_r(r,s) &= \int_0^1 t^{r-1}(1-t)^{s-1}\log(t)dt \\
&= \beta(r,s)(\psi(r)-\psi(r+s))
\end{aligned}$$

where ψ is the digamma function.

3 Probability density of k:th nearest neighbor distance

Let X be a random variable in M with a probability density function $\mu : M \to \mathbb{R}$ and $V = \{v_1, \ldots, v_n\} \subset M$ be a set of independent realizations of X. Let $q_i : M \to \mathbb{R}$, such that

$$q_i(x) = \int_{d(v_i,y)<x} \mu(y)dy.$$

That is, $q_i(x)$ is the probability that a random sample drawn from X has distance to v_i less than x. Because the upcoming results will not depend on a specific v_i, we shall drop out the reference and simply write $q(x)$ and $p(x)$. Let $v \in V$ and let $P_k : \mathbb{R} \times \mathbb{R} \to \mathbb{R} : P_k(x,y)$ be the probability for the event that, apart from v,

- There are $(k-1)$ points that have distances to v less than x.
- There are $(n-k-1)$ points that have distances to v greater than or equal to x.
- There is 1 point $w \in V$ which has distance to v less than y.

This is easily turned into a formula:

$$P_k(x,y) = C_k^* q(x)^{k-1}(1-q(x))^{n-k-1}q(y)$$

where C_k^* is a constant determined by requiring the probability density corresponding to P_k to integrate to 1. The probability that w has the distance in the range $[y, y+\Delta y]$ to v is given by:

$$\begin{aligned}
P_k(x,y+\Delta y) - P_k(x,y) &= C_k^* q(x)^{k-1}(1-q(x))^{n-k-1}\left(q(y+\Delta y)-q(y)\right) \\
&= C_k^* q(x)^{k-1}(1-q(x))^{n-k-1}q'(y_0)\Delta y
\end{aligned}$$

where the last step is by the mean value theorem and $y_0 \in [y, y + \Delta y]$. Thus:

$$\frac{P_k(x, y + \Delta y) - P_k(x, y)}{\Delta y} = C_k^* q(x)^{k-1}(1 - q(x))^{n-k-1} q'(y_0).$$

By taking the limit $\Delta y \to 0$, we get the probability density function for P_k:

$$p_k^*(x, y) = C_k^* q(x)^{k-1}(1 - q(x))^{n-k-1} q'(y).$$

By letting $f_k(x) \propto p_k^*(x, x)$, we get the probability density function for the event that the k:th nearest neighbor lies at the distance x:

$$f_k(x) = C_k q(x)^{k-1}(1 - q(x))^{n-k-1} q'(x).$$

The normalization constant is determined as follows:

$$\begin{aligned}
\int_0^\infty f_k(x)dx &= \int_0^\infty C_k q(x)^{k-1}(1 - q(x))^{n-k-1} q'(x)dx \\
&= C_k \int_0^1 t^{k-1}(1 - t)^{n-k-1} dt \\
&= C_k \beta(k, n - k) \\
&= 1
\end{aligned}$$

Thus $C_k = \frac{1}{\beta(k,n-k)}$

4 Entropy estimation using k:th nearest neighbor distances

In this section we shall derive the Kozachenko-Leonenko estimator for Shannon differential entropy.

4.1 Expected logarithm of probability mass in a k-nn ball

We would like to compute the expected value $E(q)$ of the probability mass under the k:th nearest neighbor ball centered on a point v_i. However, this quantity seems to resist analytical solution, and for this reason we instead

compute $E(\log(q))$. This is done as follows:

$$\begin{aligned} E(\log(q)) &= \int_0^\infty f(x)\log(q(x))dx \\ &= \int_0^\infty C_k q(x)^{k-1}(1-q(x))^{n-k-1}q'(x)\log(q(x))dx \\ &= C_k \int_0^1 t^{k-1}(1-t)^{n-k-1}\log(t)dt \\ &= C_k\beta(k,n-k)(\psi(k)-\psi(n)) \\ &= \psi(k)-\psi(n) \end{aligned}$$

4.2 Kozachenko-Leonenko differential entropy estimator

Let X be a random variable in M with a probability density function $\mu : M \to \mathbb{R}$. The *Shannon differential entropy of* X is given by:

$$H(X) = -\int_{x\in M} \mu(x)\log(\mu(x))dx$$

We would like to compute $H(X)$. Assume that we do not know μ, or that the computation is otherwise impractical. Then we have to estimate differential entropy from a set $(x_1,\ldots,x_n)$ of realizations of X. One such estimator is given by:

$$\hat{H}(X) = -\frac{1}{n}\sum_{t=1}^{n}\log(\hat{\mu}(x(t)))$$

Let $\epsilon_t = d(x_t,\hat{x}_t)$, where $\hat{x}_t$ is the k_t:th nearest neighbor of x_t. Let $V(t) = m(B(x_t,\epsilon_t))$, i.e. the volume of the k:th nearest neighbor ball. If we assume that the probability distribution inside the k:th nearest neighbor ball is uniform, then by the previous section we can approximate its contained logarithmic probability mass by its expectation:

$$\begin{aligned} \log(\hat{\mu}(x)V(t)) &= \psi(k_t)-\psi(n) \\ &\Leftrightarrow \\ \log(\hat{\mu}(x))+\log(V(t)) &= \psi(k_t)-\psi(n) \\ &\Leftrightarrow \\ \log(\hat{\mu}(x)) &= \psi(k_t)-\psi(n)-\log(V(t)) \end{aligned}$$

And the estimator for differential entropy becomes:

$$\begin{aligned}
\hat{H}(X) &= -\frac{1}{n}\sum_{t=1}^{n}\log(\hat{\mu}(x_t)) \\
&= -\frac{1}{n}\sum_{t=1}^{n}\left[\psi(k_t) - \psi(n) - \log(V(t))\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\log(V(t)) - \frac{1}{n}\sum_{t=1}^{n}\left[\psi(k_t) - \psi(n)\right] \\
&= \hat{H}_b(X) - \hat{H}_a(X)
\end{aligned}$$

where

$$\begin{aligned}
\hat{H}_a(X) &= \frac{1}{n}\sum_{t=1}^{n}\left[\psi(k_t) - \psi(n)\right] \\
\hat{H}_b(X) &= \frac{1}{n}\sum_{t=1}^{n}\log(V(t))
\end{aligned}$$

5 Entropy combination

Let $X = (X_1, \ldots, X_m)$ be a random variable in M. An *entropy combination* is defined by

$$I(X) = -H(X) + \sum_{i=1}^{p} s_i H(X_{L_i})$$

where

1. $L_i = \{l_1^i, \ldots, l_p^i\} \subset [1, m]$
2. $s_i \in \{-1, 1\}$
3. $X_{L_i} = (X_{l_1}, \ldots, X_{l_p})$
4. $\sum_{i=1}^{p} s_i \chi_{L_i} = \chi_{[1,m]}$

We directly substitute the differential entropy estimator from the previous section into the definition of entropy combination:

$$\begin{aligned}
\hat{I}(X) &= -\hat{H}(X) + \sum_{i=1}^{p} s_i \hat{H}(X_{L_i}) \\
&= -\left(\hat{H}_b(X) - \hat{H}_a(X)\right) + \sum_{i=1}^{p} s_i \left(\hat{H}_b(X_{L_i}) - \hat{H}_a(X_{L_i})\right) \\
&= -\left(\hat{H}_b(X) - \sum_{i=1}^{p} s_i \hat{H}_b(X_{L_i})\right) + \left(\hat{H}_a(X) - \sum_{i=1}^{p} s_i \hat{H}_a(X_{L_i})\right)
\end{aligned}$$

where we will choose the k's as follows. A fixed k is used in the estimation of $\hat{H}(X)$, and if ϵ_t is the k:th nearest neighbor distance for x_t, then $k_t^{L_i}$ is the number of points whose projections to M_{L_i} have a distance to (the projection of) x_t *less than* ϵ_t (in M_{L_i}). This is an approximation since none of the projections of the points to a marginal space need lie on the surface of the marginal ϵ_t-ball. However, this approximation combined with the definition of entropy combination allows us to cancel the $\hat{H}_b$ terms. The claim is:

$$\begin{aligned}
\hat{H}_b(X) - \sum_{i=1}^{p} s_i \hat{H}_b(X_{L_i}) &= 0 \\
&\Leftrightarrow \\
\sum_{i=1}^{p} s_i \hat{H}_b(X_{L_i}) &= \hat{H}_b(X) \\
&\Leftrightarrow \\
\frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{p} s_i \log\left(V_{L_i}(t)\right) &= \frac{1}{n} \sum_{t=1}^{n} \log(V(t)) \\
&\Leftrightarrow \\
\sum_{i=1}^{p} s_i \log\left(V_{L_i}(t)\right) &= \log(V(t))
\end{aligned}$$

In the following remember that the metric in V_{L_i} is the maximum of the involved marginal metrics, resulting in a separable expression for the volume

of a ball. The proof is then as follows:

$$\begin{aligned}
\sum_{i=1}^{p} s_i \log\left(V_{L_i}(t)\right) &= \sum_{i=1}^{p} s_i \log\left(\prod_{j\in L_i} V_j(t)\right) \\
&= \sum_{i=1}^{p}\sum_{j\in Li} s_i \log\left(V_j(t)\right) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{m} s_i \chi_{L_i}(j) \log(V_j(t)) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{p} s_i \chi_{L_i}(j) \log(V_j(t)) \\
&= \sum_{j=1}^{m} \log(V_j(t)) \sum_{i=1}^{p} s_i \chi_{L_i}(j) \\
&= \sum_{j=1}^{m} \log(V_j(t)) \\
&= \log\left(\prod_{j=1}^{m} V_j(t)\right) \\
&= \log(V(t))
\end{aligned}$$

Thus we have

$$\begin{aligned}
\hat{I}(X) &= \hat{H}_a(X) - \sum_{i=1}^{p} s_i \hat{H}_a(X_{L_i}) \\
&= \psi(k) - \psi(n) - \sum_{i=1}^{p} s_i \left(\frac{1}{n}\sum_{t=1}^{n}\left[\psi(k_t^{L_i}) - \psi(n)\right]\right)
\end{aligned}$$

Note: to guarantee that the approximation for k's holds at least for one marginal space, one should use the maximum metric everywhere, i.e. $d_i(x,y) = \max_j(|x_j - y_j|)$.

6 Examples

6.1 Mutual information

$$\begin{aligned} M &= (M_1, M_2) = (X, Y) \\ p &= 2 \\ L_1 &= \{1\} = x \\ s_1 &= 1 \\ L_2 &= \{2\} = y \\ s_2 &= 1 \end{aligned}$$

$$I(X,Y) = -H(X,Y) + H(X) + H(Y)$$
$$\hat{I}(X,Y) = \psi(k) + \psi(n) - \langle \psi(k_t^x) + \psi(k_t^y) \rangle$$

6.2 Partial mutual information

$$\begin{aligned} M &= (M_1, M_2, M_3) = (X, Z, Y) \\ p &= 3 \\ L_1 &= \{1, 2\} = xz \\ s_1 &= 1 \\ L_2 &= \{2, 3\} = zy \\ s_2 &= 1 \\ L_3 &= \{2\} = z \\ s_3 &= -1 \end{aligned}$$

$$I(X,Z,Y) = -H(X,Z,Y) + H(X,Z) + H(Z,Y) - H(Z)$$
$$\hat{I}(X) = \psi(k) - \langle \psi(k_t^{xz}) + \psi(k_t^{zy}) - \psi(k_t^z) \rangle$$

6.3 Total correlation

$$\begin{aligned} M &= (M_1, \ldots, M_m) = (X_1, \ldots, X_m) \\ p &= m \\ L_i &= \{i\} \\ s_i &= 1 \end{aligned}$$

$$I(X_1,\ldots,X_m) = -H(X_1,\ldots,X_m) + \sum_{i=1}^{m} H(X_i)$$

$$\hat{I}(X_1,\ldots,X_m) = \psi(k) + (m-1)\psi(n) - \left\langle \sum_{i=1}^{m} \psi(k_t^{L_i}) \right\rangle$$

6.4 Transfer entropy

$$\begin{aligned} M &= (M_1, M_2, M_3) = (W, X, Y) \\ p &= 3 \\ L_1 &= \{1,2\} = wx \\ s_1 &= 1 \\ L_2 &= \{2,3\} = xy \\ s_2 &= 1 \\ L_3 &= \{2\} = x \\ s_3 &= -1 \end{aligned}$$

$$I(W,X,Y) = -H(W,X,Y) + H(W,X) + H(X,Y) - H(X)$$
$$\hat{I}(W,X,Y) = \psi(k) - \langle \psi(k_t^{wx}) + \psi(k_t^{xy}) - \psi(k_t^{x}) \rangle$$

6.5 Partial transfer entropy

$$\begin{aligned} M &= (M_1, M_2, M_3, M_4) = (W, X, Z, Y) \\ p &= 3 \\ L_1 &= \{1,2,3\} = wxz \\ s_1 &= 1 \\ L_2 &= \{2,3,4\} = xzy \\ s_2 &= 1 \\ L_3 &= \{2,3\} = xz \\ s_3 &= -1 \end{aligned}$$

$$I(W,X,Z,Y) = -H(W,X,Z,Y) + H(W,X,Z) + H(X,Z,Y) - H(X,Z)$$
$$\hat{I}(W,X,Z,Y) = \psi(k) - \langle \psi(k_t^{wxz}) + \psi(k_t^{xzy}) - \psi(k_t^{xz}) \rangle$$

References

[1] Alexander Kraskov. Synchronization and interdependence measures and their applications to the electroencephalogram of epilepsy patients and clustering of data. *PhD thesis*, 2004.