Author date: 2021-06-29

Variance vs entropy of one-hot vectors

E. Castedo Ellerman (castedo@castedo.com)

Abstract

DOCUMENT TYPE: Open Study Answer

QUESTION: What is the relationship between the variance and entropy of independent one-hot vectors?

This document proves inequalities relating variance, collision entropy and Shannon entropy of sequences of independent one-hot vectors.

Introduction

Both variance and entropy are measures of uncertainty. Variance assumes values vary as points in a space separated by distances. In this document, the variance of a random vector refers to the variance of the distance from its mean (sum of the variances of each component).

Random one-hot vectors are a convenient spacial representation for categorical random variables. A one-hot vector has all components equal to 0 except one component that equals 1. This representation has been used in genetics [1]. For genetic loci with only two alleles, a one-hot vector has two redundant components. "Half" of such one-hot vectors are typically used in genetics (e.g. [2] p.40, [3], [4]). The variance of the "half one-hot vector" is exactly half the variance of its full one-hot vector.

Main Result

Given N independent random one-hot vectors X_1, X_2, ..., X_N, define

$$X_* = X_1 \times X_2 \times \cdots \times X_N$$

as the Cartesian product.

The variance of X_* can be adjusted to form a lower bound to the collision entropy, $\mathrm{H}_2(X_*)$, and Shannon entropy, $\mathrm{H}(X_*)$:

$$-\log_2\left(1 - \frac{\mathrm{Var}(X_*)}{N}\right) \leq \frac{\mathrm{H}_2(X_*)}{N} \leq \frac{\mathrm{H}(X_*)}{N}$$

If every X_i takes only two equally likely values, then the lower bounds reach equality:

$$-\log_2\left(1 - \frac{\mathrm{Var}(X_*)}{N}\right) = \frac{\mathrm{H}_2(X_*)}{N} = \frac{\mathrm{H}(X_*)}{N} = 1$$

Proof

Let M_i be length of X_i (the number of categorical values represented by X_i). Let $p_{i,j}$ represent the probability of X_i taking the j-th categorical value.

For every $1 \leq i \leq N$,

$$\sum_{j=1}^{M_i} p_{i,j} = 1$$

The expectation and variance of the i-th one-hot vector X_i is

$$\begin{aligned}
\mathrm{E}(X_i) &= (\ p_{i,1}\ ,\ p_{i,2}\ ,\ \ldots\ ,\ p_{i,M_i}\ ) \\
\mathrm{Var}(X_i) &= \sum_{j=1}^{M_i} p_{i,j} \left[(1 - p_{i,j})^2 + \sum_{k \neq j} (0 - p_{i,k})^2 \right] \\
&= \sum_{j=1}^{M_i} p_{i,j} \left[1 - 2p_{i,j} + \sum_{k=1}^{M_i} p_{i,k}^2 \right] \\
&= 1 - 2 \sum_{j=1}^{M_i} p_{i,j}^2 + \sum_{k=1}^{M_i} p_{i,k}^2 \\
&= 1 - \sum_{j=1}^{M_i} p_{i,j}^2
\end{aligned}$$

Thus the variance of X_i equals the probability of two independent samples from X_i being distinct. This probability of distinction has been called logical entropy [5].

The complement

$$1 - \mathrm{Var}(X_i) = \sum_{j=1}^{M_i} p_{i,j}^2$$

is the chance of repetition, which is expected probability. Taking the negative log gives Rényi entropy of order 2, also called collision entropy:

$$-\log_2 (1 - \mathrm{Var}(X_i)) = -\log_2 \left(\sum_{j=1}^{M_i} p_{i,j}^2 \right) = \mathrm{H}_2(X_i)$$

Since negative log is a concave function, the negative log of expected probability (collision entropy), is a lower bound to the expected negative log of probability (Shannon entropy) by Jensen's inequality:

$$\mathrm{H}_2(X_i) = -\log_2 \left(\sum_{j=1}^{M_i} p_{i,j}^2 \right) \leq \sum_{j=1}^{M_i} p_{i,j} (-\log_2 p_{i,j}) = \mathrm{H}(X_i)$$

The total variance, can be adjusted to equal the average probability of one-hot vector repetition (per one-hot vector):

$$1 - \frac{\mathrm{Var}(X_*)}{N} = 1 - \frac{1}{N} \sum_{i=1}^{N} \mathrm{Var}(X_i) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} p_{i,j}^2$$

Negative log with Jensen's inequality can then establish yet another lower bound:

$$-\log_2\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_i} p_{i,j}^2\right) \leq \frac{1}{N}\sum_{i=1}^{N}\left(-\log_2\sum_{j=1}^{M_i} p_{i,j}^2\right) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{H}_2(X_i)$$

Collision and Shannon entropy are additive for independent variables. Putting everything together we get

$$-\log_2\left(1-\frac{\mathrm{Var}(X_*)}{N}\right) \;\leq\; \frac{\mathrm{H}_2(X_*)}{N} \;\leq\; \frac{\mathrm{H}(X_*)}{N}$$

References

1. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science. 1978;201: 786–792. doi:10.1126/science.356262
2. Weir BS. Genetic data analysis II: Methods for discrete population genetic data. Sunderland, Mass: Sinauer Associates; 1996.
3. Weir BS, Hill WG. Estimating F-Statistics. Annual Review of Genetics. 2002;36: 721–750. doi:10.1146/annurev.genet.36.050802.093940
4. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genetics. 2006;2: e190–. doi:10.1371/journal.pgen.0020190
5. Ellerman D. Logical information theory: New logical foundations for information theory. Logic Journal of the IGPL. 2017;25: 806–835. doi:10.1093/jigpal/jzx022