THE VON NEUMANN ENTROPY FOR DATA COMPRESSION

Aidan Rocke
aidanrocke@gmail.com

September 8, 2021

ABSTRACT

This article introduces an efficient information-theoretic method for dimensionality reduction in Hilbert Spaces. The efficacy of this method for choosing informative dimensions relies upon the fact that the Von Neumann entropy of a normalised Pearson Correlation matrix is well-defined.

1 The Von Neumann entropy of a Pearson Correlation matrix

Let's suppose we have an ergodic dynamical system with N observables $x_i(t) \in \mathbb{R}$ which are sampled using a sequence of n measurements so we have a dataset $X \in \mathbb{R}^{N \times n}$. Given X, we may compute the statistics $X_i = x_i - \langle x_i \rangle$ and $\sigma_i^2 = \langle X_i^2 \rangle$ which allows us to define the Pearson Correlation Matrix with entries:

$$R_{i,j} = \frac{X_i \cdot X_j}{\sigma_i \cdot \sigma_j} \tag{1}$$

Given $R \in \mathbb{R}^{N \times N}$, we may define the density matrix $\rho = \frac{R}{N}$ which is positive semi-definite, Hermitian and has unit-trace. Thus, we may calculate the entropy of R using the Von Neumann entropy:

$$S(\rho) = -\text{tr}(\rho \cdot \ln \rho) = -\sum_{i=1}^{N} \lambda_i \cdot \ln \lambda_i \tag{2}$$

where λ_i are the eigenvalues of ρ.

2 Using the Von Neumann entropy for dimensionality reduction

If N is large, in order to compress the dataset X so that we keep the dimensions that contain 95% of the statistical information it is sufficient to find the discrete subset $S \subset [1, N]$ that maximises:

$$-\sum_{i \in S} \lambda_i \cdot \ln \lambda_i \tag{3}$$

subject to the constraint $\frac{-\sum_{i \in S} \lambda_i \cdot \ln \lambda_i}{-\sum_{i=1}^{N} \lambda_i \cdot \ln \lambda_i} \leq \frac{95}{100}$ which may be done using sorting algorithms such as Quick Sort.

3 Theory

Assuming the Manifold Hypothesis is true [8], the intrinsic dimension of an ergodic dynamical system evolving in a Hilbert Space is much smaller than its ambient dimension. For datasets where this assumption is valid, the cardinality $|S|$ as calculated in (3) provides us with an upper-bound on its intrinsic dimension.

References

[1] von Neumann, John (1932). Mathematische Grundlagen der Quantenmechanik (Mathematical Foundations of Quantum Mechanics) Princeton University Press., . ISBN 978-0-691-02893-4.

[2] H. Felippe et al. The von Neumann entropy for the Pearson correlation matrix: A test of the entropic brain hypothesis. Arxiv. 2021.

[3] E.T. Jaynes. Information Theory and Statistical Mechanics. The Physical Review. 1957.

[4] Sharpee, Tatyana, Nicole C. Rust, and William Bialek. Maximally informative dimensions: analyzing neural responses to natural signals. Advances in Neural Information Processing Systems (2003): 277-284.

[5] Edward Witten. A Mini-Introduction To Information Theory. Arxiv. 2019.

[6] Karl Friston. The free-energy principle: a rough guide to the brain? Cell Press. 2009.

[7] Aidan Rocke (https://mathoverflow.net/users/56328/aidan-rocke), Physical interpretation of the Manifold Hypothesis, URL (version: 2020-01-28): https://mathoverflow.net/q/351368