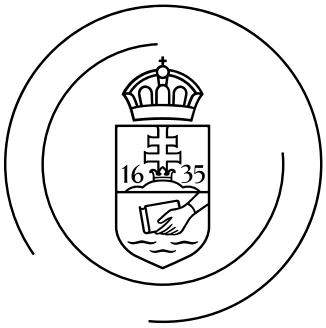
Math project

Entropy estimation and high entropy projections

Emma Lukács

Supervisor: Adrián Csiszárik



EÖTVÖS LORÁND
UNIVERSITY | BUDAPEST

Applied Mathematics MSc
2023/2

1. Introduction

The information theory paradigm, rooted in Shannon's foundational work from the 1940s [6], has gained significant traction in Machine Learning and Neural Networks. Self-supervised learning, which involves models predicting one part of the input from another, reflects principles akin to entropy maximization. Despite its historical significance, several fundamental questions persist in the field. One major challenge is the usage of information entropy in real-world scenarios due to the absence of underlying Probability Density Functions (PDFs), leaving only observed data. This obstacle necessitates accurate entropy estimation solely from observed data, driving the need for flexible, non-parametric methods like the k-nearest neighbor (kNN) approach pioneered by Kozachenko and Leonenko [4]. However, the classical kNN estimator exhibits bias, particularly in higher dimensions [5]. A significant part of my investigation involved identifying and applying the most effective projection techniques to reduce dimensionality while preserving essential information. This initiative is driven by a comparative analysis of various projection methods, including Principal Component Analysis (PCA) and variance-based techniques. These methods are well-regarded for capturing the global structure of data. By integrating these with the locality-sensitive kNN approach, I aim to refine our methods for entropy estimation and improve our understanding of complex data distributions in reduced-dimensional spaces. This comprehensive approach is intended to mitigate the biases associated with high-dimensional kNN entropy estimation and enhance the overall accuracy of entropy measures.

2. PCA and the Kozachenko-Leonenko estimate

2.1. Definition (Principal Component Analysis)**.** According to Jolliffe [3] PCA is defined as an orthogonal linear transformation on a real inner product space that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.
Consider an $n \times p$ data matrix, $\mathbf{X}$, with column-wise zero empirical mean.
The transformation is defined by a set of size l of p-dimensional vectors of weights or coefficients $\mathbf{w}(k) = (w_{1(k)}, \dots, w_{p(k)})^\top$ that map each row vector $\mathbf{X}(i) = (x_{1(i)}, \dots, x_{p(i)})$ of $\mathbf{X}$ to a new vector of principal component scores $\mathbf{t}(i) = (t_{1(i)}, \dots, t_{l(i)})$, given by

$$t_k(i) = \mathbf{X}(i) \cdot \mathbf{w}(k) \quad \text{for } i = 1, \dots, n \text{ and } k = 1, \dots, l,$$

in such a way that the individual variables $t_1, \dots, t_l$ of $\mathbf{t}$ considered over the data set successively inherit the maximum possible variance from $\mathbf{X}$, with each coefficient vector $\mathbf{w}$ constrained to be a unit vector (where l is usually selected to be strictly less than p to reduce dimensionality).
The above may equivalently be written in matrix form as

$$\mathbf{T} = \mathbf{X}\mathbf{W},$$

where $T_{ik} = t_k(i)$, $X_{ij} = x_j(i)$, and $W_{jk} = w_{j(k)}$.
First component
In order to maximize variance, the first weight vector $\mathbf{w}(1)$ thus has to satisfy

$$\mathbf{w}(1) = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (\mathbf{X}(i) \cdot \mathbf{w})^2 \right\} \tag{1}$$

Equivalently, writing this in matrix form gives

$$\mathbf{w}(1) = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\mathbf{X}\mathbf{w}\|^2 \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} \right\}$$

2.2. Definition (Entropy)**.** Let X be a discrete random variable with probability mass function $P_X(x)$, $x \in \mathcal{X}$. The *entropy* (or *Shannon entropy*) of X is

$$H(X) = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \tag{2}$$

$$= \int_{\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}\, dx. \tag{3}$$

2.3. Definition (k-Nearest Neighbour Kozachenko-Leonenko estimator)**.** According to definition introduced in the paper written by Ao and Li [1]. Let $x_1, x_2, \ldots, x_n$ $(n \geq 3)$ be i.i.d. random variables with density f on $\mathbb{R}^d$. Let us indentify the k-nearest neighbors (in terms of the p-norm distance) for each x_i and define the smallest closed ball covering them as:

$$B(x_i, \frac{\varepsilon_i}{2}) = \{x \in \mathbb{R}^d | \|x - x_i\|_p \leq \frac{\varepsilon}{2}\},$$

where ε is twice the distance of x_i and its k-th nearest neighbour, and the mass of $B(x_i, \frac{\varepsilon_i}{2})$ is:

$$q_i(\varepsilon_i) = \int_{x\in B(x_i, \frac{\varepsilon_i}{2})} P_X(x)dx \Rightarrow \mathbb{E}(\log(q_i)) = \psi(k) - \psi(N),$$

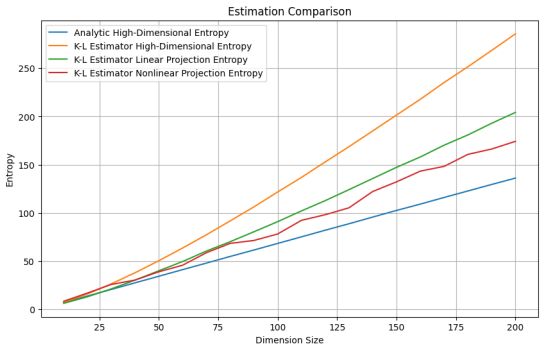
where $\psi(N)$ is equal to $\frac{\Gamma'(x)}{\Gamma(x)}$ with $\Gamma(x)$ being the Gamma function. The main assumption of the KL estimation is that the density is constant within the unit ball approximated by $q_i(\varepsilon_i) \approx c_d \varepsilon_i^d P_X(x_i)$, where d is the dimension of X and c_d is given by $\frac{\Gamma(1+\frac{1}{p})^d}{\Gamma(1+\frac{d}{p})}$, which is the volume of the d-dimensional unit ball according to the given p-norm. This yields the final KL-estimator formula:

$$\hat{H}_{KL} = \psi(k) + \psi(N) + \log c_d + \frac{d}{N}\sum_{i=1}^{N} \log(\varepsilon_i). \tag{4}$$

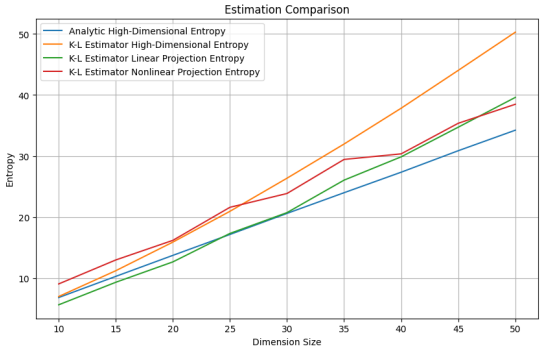
3. Dimensionality Reduction using Autoencoders

I explored dimensional reduction using autoencoders, systems designed to encode high-dimensional data $\mathbf{x} \in \mathbb{R}^{\text{input dim}}$ into a more compact latent space $\mathbf{z} \in \mathbb{R}^{\text{latent dim}}$ through an encoder function f_{enc}, and then reconstruct it via a decoder function f_{dec}, where $\mathbf{z} = f_{\text{enc}}(\mathbf{x})$ and $\mathbf{x}' = f_{\text{dec}}(\mathbf{z})$ [2]. This investigation initially centered on the interplay between such projections and variance-based dimensionality reduction methods. Two specific types of autoencoders were examined: a **Nonlinear Autoencoder**, which employs multiple hidden layers with ReLU activations, and a **Linear Autoencoder** that uses a single linear transformation for both encoding and decoding. The models were trained using the Mean Squared Error (MSE) loss and the Adam optimizer. To generate suitable data, I first produced multidimensional normal distributions with an emphasis on clustering around a hyperplane by scaling select dimensions, and also employed Cholesky decomposition to create sets of decorrelated Gaussian data.

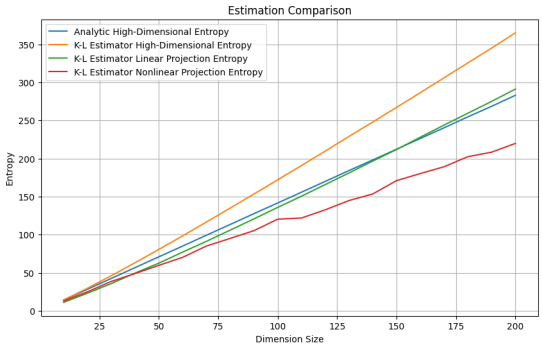
I assessed entropy in a high-dimensional dataset and its lower-dimensional projections derived from two autoencoder models, comparing these against both analytical methods and the KL-estimate. The visualizations indicated that, generally, the nonlinear autoencoder performed best at entropy estimation in lower-dimensional spaces, as shown in Figure 1. In cases where data closely resembled Gaussian distributions, the linear autoencoder's projections—akin to PCA results—yielded entropy estimates that closely matched the analytical values, particularly in dimensions exceeding 60. This observation highlights the effectiveness of variance-based methods in handling Gaussian-like data, and the capability of linear autoencoders to produce projections that are not only comparable to PCA but also enhance entropy estimation. These findings encouraged further exploration into the relationship between these projection techniques and more complex data scenarios, aiming to understand better how these models preserve critical information for accurate entropy assessment.



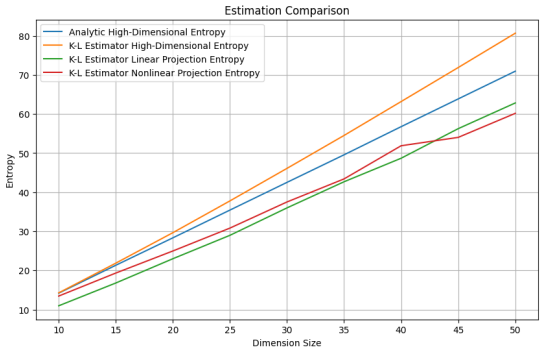
((a)) First data - full dimension range



((b)) First data - lower dimension range



((c)) Second data - full dimension range



((d)) Second data - lower dimension range

Figure 1. Comparison of entropy estimation across dimension sizes

The plot color scheme is coherent: blue indicates the analytic entropy, orange represents the original dimensional KL-estimate, green stands for the linear projection KL-estimate, and red signifies the nonlinear projection. The 'first data' pertains to the initial data generation process, while the 'second data' corresponds to the decorrelated Gaussian data.

4. Projections preserving maximal entropy

I aimed to delve deeper into advanced and appropriate projection methods that would be more beneficial for this specific area. Autoencoders, through their loss functions, capture and focus on PCA-like and variance-based relationships. However, I now seek a change of perspective by exploring the mathematical foundations of Principal Component Analysis (PCA) and variance against the entropy estimates derived from the Kozachenko-Leonenko (KL) method.

PCA operates under the assumption that the data is Gaussian distributed and it is effective in reducing dimensionality while preserving as much variance as possible, capturing the global structure of the data. In contrast, the KL method estimates entropy by considering the distances to the nearest neighbors within the data. This method does not rely on a specific distribution but instead generalizes to the geometry of the data within a unit ball (or a unit circle in the 2D case). By focusing on nearest neighbor relations, the KL method captures the local structure and variability of the data, providing a more accurate entropy estimate for complex, non-Gaussian distributions. This shift in perspective highlights the trade-off between PCA and the KL method. By integrating these insights, I hope to develop more effective projection methods for analyzing lower-dimensional spaces.

.

4.1. Definition (Maximum Sliced Entropy)**.** The formal definition of max sliced entropy was introduced by Dor Tsur, Ziv Goldfeld, and Kristjan Greenewald [7]. Given a random variable X with distribution μ_X in $\mathbb{R}^d$, the *k-dimensional Maximum Sliced Entropy* (MSE) of X, denoted as $sh_k(X)$, is defined by:

$$sh_k(X) = \sup_{A \in St(k,d)} h(A^T X),$$

where $St(k, d)$ represents the Stiefel manifold of all $d \times k$ matrices with orthonormal columns, and h

denotes the differential entropy.

Maximum Sliced Entropy (MSE) is instrumental in capturing the most informative features of high-dimensional data. MSE, by projecting the data onto lower-dimensional subspaces, aims to maximize the entropy across various data slices, thus ensuring that the projections retain significant diversity and information content of the original dataset.

4.1. PCA and Max Sliced Entropy

Based on the paper by Tsur, Goldfeld and Greenewald [7] an important connection for Gaussian data was also formally revealed.

Proposition. *(Equivalence of Max-Sliced Entropy and PCA)*
Let $X \sim N(m, \Sigma)$ with $m = 0$ and $\Sigma \in \mathbb{R}^{d\times d}$ being full-rank. The equivalence between the max-sliced entropy and PCA is established under the assumption that the k-dimensional PCA for Σ is given by the optimization problem:

$$\sup_{A\in St(k,d)} \operatorname{tr}(A^T \Sigma A),$$

where A_{PCA} is the matrix that contains the first k eigenvectors of Σ, which correspond to its largest k eigenvalues.

Proof:
Define the max-sliced entropy of X, $sh_k(X)$, as:

$$sh_k(X) = \sup_{A\in St(k,d)} h(A^T X) = \sup_{A\in St(k,d)} \frac{1}{2} \log\left((2\pi e)^k \det(A^T \Sigma A)\right),$$

which simplifies further using the properties of the determinant and eigenvalues:

$$= \sup_{A\in St(k,d)} \frac{1}{2} \sum_{i=1}^{k} \log(2\pi e \lambda_i(A^T \Sigma A)) = \frac{1}{2} \sum_{i=1}^{k} \log(2\pi e \lambda_i(\Sigma)),$$

where the second equality is derived from the differential entropy of a k-dimensional Gaussian random vector and the last equality holds due to the eigenvalue relations:

$$\lambda_{d-k+i}(\Sigma) \leq \lambda_i(A^T \Sigma A) \leq \lambda_i(\Sigma) \quad \text{for } i = 1, \ldots, k.$$

This is justified by the Poincaré separation theorem which implies the interlacing of the eigenvalues of Σ and $A^T \Sigma A$. The monotonicity of the logarithm function then concludes the proof, showing the equivalence between the maximization of the entropy $h(ATX)$ via PCA and the optimization of the variance captured by the top k components.

5. A new approach for estimating the direction

Recognizing that PCA is an effective estimator of Max Sliced Entropy in data that exhibits Gaussian characteristics, I wanted to construct an algorithm that strategically segments the dataset to enhance its Gaussian-likeness using Gaussian Mixture Models (GMMs). Furthermore, it incorporates a component weighing system that captures a balance between local and global properties of the data. This weighing system remains intentionally generalized to accommodate future explorations and refinements. To ensure consistency in the direction of the principal component vectors, I transformed each PCA vector into a canonical form. Specifically, if the first component of the PCA vector $\mathbf{w}_k$ was negative, I multiplied the entire vector by -1. This can be expressed as:

$$\mathbf{w}_k^{\text{canonical}} = \begin{cases} \mathbf{w}_k & \text{if } w_{k1} \geq 0 \\ -\mathbf{w}_k & \text{if } w_{k1} < 0 \end{cases}$$

1. Algorithm Maximal Entropy and Cluster-based PCA Direction Estimation

Require: Dataset $\mathbf{D}$, angle range 0 to 360 degrees
Step 1: Estimate Maximal Entropy Direction
for each angle θ from 0 to 360 degrees **do**
Project dataset $\mathbf{D}$ onto angle θ
Compute the entropy $H(\theta)$ using histogram-based estimation
end for
Identify the angle θ^* that maximizes $H(\theta)$
Maximal entropy direction θ^*
Step 2: Cluster Analysis using GMMs
Fit GMM to dataset $\mathbf{D}$
Determine optimal number of clusters K using BIC
Classify data into K clusters
Step 3: PCA for Each Cluster
for each cluster k in 1 to K **do**
Extract data points $\mathbf{D}_k$ belonging to cluster k
Perform PCA on $\mathbf{D}_k$
Obtain principal component directions $\mathbf{w}_k$
end for
Step 4: Compute Generalized Weighted Average of PCA Directions
Introduce a generalized weighting scheme for further analysis
Calculate a weighted average PCA direction based on the selected scheme
Evaluate and compare the results
Output: Maximal entropy direction generalized weighted average PCA direction

In further exploring the structure of cluster connections, I employed a novel approach by representing each cluster by a single point in a higher-dimensional space. This representation allowed me to analyze the relationships between clusters by projecting these points onto a line. The primary focus was to examine the entropy of these projections.

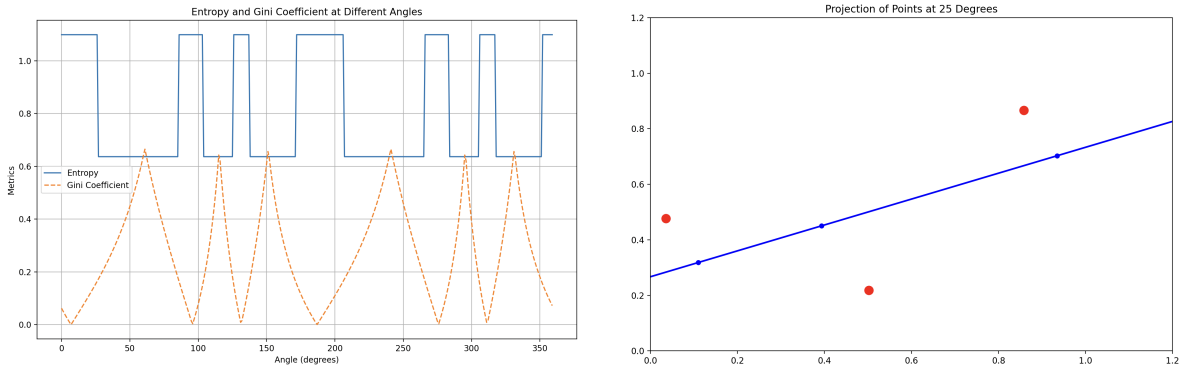


Figure 2. Structure of entropies at different angle projections
The first plot shows the entropy and Gini coefficient of projections at different angles and helps identify angles that maximize or minimize the randomness and inequality of the point distribution, while the second plot provides a visual representation of the point distribution at particular "peak" angle.

The findings revealed that the uniformity of the nearest neighbor distances, quantified using Gini coefficients, was a key metric in capturing the maximization of entropy. Significantly, I found that peaks in the projection entropy were associated with two main properties: one is the uniformity of the nearest neighbor distances, and the other is the length of these distances. These insights suggest that not only the dispersion of points impacts entropy but also the relative distances among them. This dual aspect underscores the importance of spatial relationships in the data, influencing the entropy outcomes and providing a deeper understanding of how entropy behaves in response to changes in cluster configurations. Moving forward, I aim to incorporate this knowledge into the development of a more sophisticated algorithm. The goal is to integrate the understanding of uniformity and distance measures into the weighing method used within the algorithm.

Bibliography

[1] Ziqiao Ao and Jinglai Li. Entropy estimation via uniformization. *Artificial Intelligence*, 322:103954, 2023.

[2] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021.

[3] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York, Inc., New York, 2002.

[4] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf.*, 23(2):9–16, 1987. Translated from: Problems Inform. Transmission, 23(2):95–101, 1987.

[5] Chien Lu and Jaakko Peltonen. Enhancing nearest neighbor based entropy estimator for high dimensional distributions via bootstrapping local ellipsoid. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:5013–5020, 04 2020.

[6] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[7] Dor Tsur, Ziv Goldfeld, and Kristjan Greenewald. Max-sliced mutual information. *arXiv preprint arXiv:2309.16200*, 2023.